

AN INTERNAL PILOT STUDY WITH INTERIM ANALYSIS FOR GAUSSIAN LINEAR MODELS

John A. Kairalla

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics, School of Public Health.

Chapel Hill
2007

Approved by:

Chairman: Keith E. Muller

Co-Chairman: Christopher S. Coffey

Reader: Lisa M. LaVange

Reader: Todd A. Schwartz

Reader: Annelies Van Rie

ABSTRACT

JOHN A. KAIRALLA. An Internal Pilot Study with Interim Analysis
for Gaussian Linear Models.
(Under the direction of Keith E. Muller and Christopher S. Coffey)

Misspecification of a nuisance parameter can lead to study power far from the desired level. Internal pilots for Gaussian data protect study power by allowing sample size re-estimation based on an interim power analysis using a revised estimate of the variance parameter, but without any data analysis. In order to reduce study time and cost, researchers and sponsors of studies often desire early decision possibilities that the internal pilot design lacks but that group sequential methods allow. Combining early stopping rules with internal pilot methods would increase study flexibility, scope, and efficiency for general linear models.

An internal pilot with an interim analysis (IPIA) design for Gaussian linear models is introduced and defined. The design allows for early stopping for efficacy and futility while also re-estimating sample size needs based on an interim variance estimate. In order for accurate study planning in small samples, exact theory is derived for both the one or two group t test setting, as well as more complex multiple degree of freedom hypothesis tests within the general linear univariate model framework. Exact and computable forms of distributions allow accurate calculations of power, type I error rate, and expected sample size.

In general, the IPIA design maintains and controls power to the desired level and also provides sample size savings. However, it can also inflate type I error rate, especially in smaller studies. By utilizing the exact theory, planning procedures associated with the design are examined and refined to create a working method for planning sound studies. A

bounding method successfully controls the type I error rate while maintaining the benefits of the design. Explicit recommendations are detailed that achieve the combined goals of an internal pilot and a two-stage group sequential design. The results can be used during planning to create an efficient two-stage study with early stopping rules and predictable power properties, even in small samples.

ACKNOWLEDGEMENTS

I wish to thank my dissertation chairmen Keith Muller and Chris Coffey for all of their support and encouragement during the course of this research. Their active interest allowed this work to prosper despite often many miles of separation. I am very grateful to have had the opportunity to work closely and learn from both of you (including the late night theory sessions that went on long past Keith's bedtime). I would also like to thank my committee members Lisa LaVange, Todd Schwartz, and Annelies Van Rie. In addition to their unwavering support and flexibility, they each provided valuable insight and context for this work that will no doubt help my career progress. I truly hope to continue working with each of them in the years ahead.

I would also like to thank the UNC Department of Biostatistics for all of the support I have received. First, Craig Turnbull introduced me to the field as an undergraduate and prepared me for graduate research as adviser of my honor's thesis. Second, I am very grateful to NIH grant funding I have received with PIs: Ed Davis, Lloyd Edwards, Larry Kupper, and Amy Herring. Also, my GRAs with Richard Henderson and Keith Muller were incredibly valuable experiences in consulting, programming, organizing, and document preparation as well as financially supportive. Finally, the staff and students in biostatistics have been wonderful to be around and are indispensable to my graduate education.

I also need to thank my wife Ashley for her love and support. She has been so understanding of my lack of free time or energy for household duties and familial time over the last number of years. She has done a tremendous job of keeping our family close and functional despite returning to school herself and getting a nursing degree from UNC. With this time behind us, I look forward to many years of work, growth, and happiness.

TABLE OF CONTENTS

LIST OF TABLES.....	ix
LIST OF ABBREVIATIONS.....	xi
Chapter	
1. INTRODUCTION AND LITERATURE REVIEW.....	1
1.1 INTRODUCTION.....	1
1.2 LITERATURE REVIEW.....	3
1.2.1 Introduction.....	3
1.2.2 Group Sequential Methods.....	4
1.2.3 Stochastic Curtailment.....	8
1.2.4 Sample Size Re-Estimation.....	10
1.2.4.1 Introduction.....	10
1.2.4.2 Flexible Designs.....	11
1.2.4.3 Adaptive Designs.....	11
1.2.4.4 Sample Size Re-Estimations Based on Conditional Power.....	13
1.2.4.5 Internal Pilot Designs.....	14
1.2.4.6 Review of Related Topics.....	17
1.3 SUMMARY.....	21
2. INTERNAL PILOT WITH INTERIM ANALYSIS FOR SINGLE DEGREE OF FREEDOM HYPOTHESIS TESTS.....	22
2.1 INTRODUCTION.....	22
2.1.1 Motivation.....	22
2.1.2 Literature Review.....	23

2.2 THE IPIA MODEL AND PROPERTIES.....	27
2.2.1 Notation.....	27
2.2.2 The IPIA Model.....	28
2.2.3 IPIA Properties.....	32
2.3 THE IPIA PROCEDURE AND PROPERTIES.....	40
2.4 KEY ANALYTIC RESULTS FOR PROCEDURE.....	43
2.5 EXAMPLES.....	46
2.5.1 Motivation for Examples.....	46
2.5.2 Computational Methods.....	48
2.5.3 Example 2.1 Results.....	50
2.5.4 Example 2.2 Results.....	54
2.6 DISCUSSION.....	58
3. PLANNING PROCEDURES FOR AN INTERNAL PILOT WITH INTERIM ANALYSIS DESIGN.....	62
3.1 INTRODUCTION.....	62
3.1.1 Motivation.....	62
3.1.2 Literature Review.....	63
3.2 THE IPIA MODEL AND PROCEDURE.....	65
3.2.1 Notation.....	65
3.2.2 The IPIA Model.....	66
3.2.3 The General Procedure.....	68
3.3 CRITICAL VALUE SELECTION.....	68
3.3.1 Overview.....	68
3.3.2 Distributional Assumptions.....	69
3.3.3 The IPIA Bounding Method.....	70
3.4 SAMPLE SIZE SELECTION METHODS.....	71
3.4.1 Sample Size Re-Estimation.....	71

3.4.2 Interim Sample Size Selection.....	73
3.5 EXAMPLES.....	74
3.5.1 Example Motivation.....	74
3.5.2 Example Methods.....	74
3.5.3 Computational Methods.....	75
3.5.4 Example 3.1 Results.....	76
3.5.5 Example 3.2 Results.....	79
3.5.6 Interim Sample Size Selection Results.....	81
3.6 DISCUSSION.....	83
4. INTERNAL PILOT WITH INTERIM ANALYSIS FOR MULTIPLE DEGREE OF FREEDOM HYPOTHESIS TESTS.....	87
4.1 INTRODUCTION.....	87
4.1.1 Motivation.....	87
4.1.2 Literature Review.....	88
4.2 THE IPIA MODEL AND PROPERTIES.....	89
4.2.1 Notation.....	89
4.2.2 The IPIA Model.....	90
4.2.3 IPIA Properties.....	92
4.3 THE IPIA PROCEDURE AND PROPERTIES.....	101
4.4 KEY ANALYTIC RESULTS FOR PROCEDURE.....	104
4.5 AN EXAMPLE.....	107
4.5.1 Motivation for the Example.....	107
4.5.2 Computational Methods.....	110
4.5.3 Example 4.1 Results.....	111
4.6 DISCUSSION.....	116
5. SUMMARY AND FUTURE RESEARCH.....	119
5.1 SUMMARY OF ACCOMPLISHMENTS.....	119

5.1.1 Chapter 2: Internal Pilot with Interim Analysis for Single Degree of Freedom Hypothesis Tests.....	119
5.1.2 Chapter 3: Planning Procedures for an Internal Pilot with Interim Analysis Design.....	119
5.1.3 Chapter 4: Internal Pilot with Interim Analysis for Multiple Degree of Freedom Hypothesis Tests.....	120
5.2 FUTURE RESEARCH.....	121
5.2.1 Futility Bounds.....	121
5.2.2 Sample Size Re-Estimation Method.....	121
5.2.3 Selection of Interim Sample Size.....	122
5.2.4 Computation.....	123
5.2.5 Strategies for Multiple Degree of Freedom Tests.....	123
5.2.6 Generalizations to Other Settings.....	124
APPENDIX A: CHAPTER 2 PROOFS.....	126
APPENDIX B: CHAPTER 4 PROOFS.....	133
REFERENCES.....	144

LIST OF TABLES

Table

2.1	Dimensions.....	31
2.2	Parameters and constants.....	31
2.3	Internal pilot with interim analysis notation.....	32
2.4	General procedure.....	40
2.5	Two-stage designs.....	47
2.6	Simulation and calculation times (minutes).....	49
2.7	Design parameters for Example 2.1.....	50
2.8	Type I error rates $\times 100$ for Example 2.1.....	50
2.9	Power $\times 100$ for Example 2.1.....	51
2.10	$E(N_w)$ for Example 2.1: fixed, IP, and GS.....	52
2.11	$E(N_w)$ for Example 2.1: IPIA.....	53
2.12	Design Parameters for Example 2.2.....	54
2.13	Type I error rates $\times 100$ for Example 2.2.....	55
2.14	Power $\times 100$ for Example 2.2.....	56
2.15	$E(N_w)$ for Example 2.2: fixed, IP, and GS.....	57
2.16	$E(N_w)$ for Example 2.2: IPIA.....	57
3.1	General Procedure.....	68
3.2	Design parameters for Example 3.1.....	76
3.3	Type I error rates $\times 100$ for Example 3.1.....	76
3.4	Power $\times 100$ for Example 3.1.....	77
3.5	$E(N_w)$ for Example 3.1.....	78
3.6	Design parameters for Example 3.2.....	79
3.7	Type I error rates $\times 100$ for Example 3.2.....	79
3.8	Power $\times 100$ for Example 3.2.....	80
3.9	$E(N_w)$ for Example 3.2.....	80

3.10	Type I error rates $\times 100$ for $\pi \in \{0.25, 0.5, 0.75\}$	81
3.11	Power $\times 100$ for $\pi \in \{0.25, 0.5, 0.75\}$	82
3.12	$E(N_w)$ for $\pi \in \{0.25, 0.5, 0.75\}$	83
4.1	General procedure.....	101
4.2	Two-stage design.....	108
4.3	Simulation and calculation times (hours) for Ex. 4.1.....	111
4.4	Design parameters for Example 4.1.....	112
4.5	Type I error rates $\times 100$ for Example 4.1.....	112
4.6	Power $\times 100$ for Example 4.1.....	113
4.7	$E(N_w)$ for Example 4.1: fixed, IP, and GS.....	113
4.8	$E(N_w)$ for Example 4.1: IPIA.....	115

LIST OF ABBREVIATIONS

AD	Adaptive design
CP	Conditional power
GLUM	General linear univariate model
GSM	Group sequential method
IP	Internal pilot
IPIA	Internal pilot with interim analysis
SPRT	Sequential probability ratio test
SSR	Sample size re-estimation

CHAPTER 1. INTRODUCTION AND LITERATURE REVIEW

1.1 INTRODUCTION

An important aspect to consider during study planning is an appropriate sample size to detect an effect of interest for given type I error rate and power. Power in studies often depends on one or more unknown nuisance parameters. For example, in a one group t -test, let X_1, X_2, \dots be independent Gaussian observations with mean θ and unknown error variance σ^2 . The goal is to test $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$ with type I error rate α_t and power P_t at $\theta = \theta_1$. The required sample size for the test with given α_t, P_t , and θ_1 depends on σ^2 . In practice, sample size needs are usually calculated using an estimated value, σ_0^2 , taken from similar studies or earlier trials. This value, however, is often not appropriate due to characteristics such as differing populations or inadequate sample size of test trials. Study power for a Gaussian linear model is very sensitive to misspecification of the variance parameter, σ^2 .

In general, sample size re-estimation techniques have been developed as tools for adjusting the size of a study to meet its planned objectives. To ensure a correctly planned study, very few interim analyses are conducted. In fact, two-stage designs have become popular due to their practicality, effectiveness, and lack of administrative burden (Shih, 2006). Specifically, *internal pilot* (IP) designs are two-stage designs that allow sample size modification based on revised estimates of nuisance parameters without interim data analysis. For continuous data, the IP design introduced by Wittes and Brittain (1990) used the ordinary (unadjusted) test statistic and critical value for testing. This procedure is straightforward to implement, but may introduce type I error rate inflation in small samples. For a general linear univariate model (GLUM) with fixed effects and Gaussian errors, Coffey

and Muller (1999) derived the exact distribution of the IP test statistic in a computable form. The theory includes t -tests as special cases and allows for flexibility of hypotheses. The exact theory also made study planning possible for small sample designs.

Researchers and sponsors of clinical trials and other studies would also like the ability to reach early decisions when hypothesis outcomes are clear. Early decisions to stop a trials may allow more effective treatments to reach a target population quickly and can protect patients from ineffective, inefficient, or harmful treatments. Stopping early can also allow resources to be diverted to other promising research, boosting overall research efficiency. To address the need for early stopping capability in study design, study monitoring procedures such as group sequential and stochastic curtailment methods have been developed.

Combining IP designs with early stopping rules would increase study flexibility, scope, and efficiency for general linear models. In this dissertation, the exact distributions necessary for small sample internal pilots with an interim analysis at the IP stage for GLUMs with fixed effects and Gaussian errors are derived. The study design is hence referred to as the internal pilot with interim analysis (IPIA) design. In Chapter 2 of this dissertation, the model is introduced, the procedure is explained, and the necessary distributions for the IPIA design for single degree of freedom tests are derived. These study designs consist of one and two group comparisons with unknown, common variances for t -tests as well as other study designs. The necessary distributions include a computable form of the exact joint distribution of the first and second stage test statistics conditional on a final sample size. Knowledge of these forms then allow for derivation of exact forms for unconditional study power, type I error rate, and expected sample size. Examples will portray the characteristics of the IPIA design and compare them with some other common designs.

The procedure introduced in Chapter 2 has various general design factors including the methods for selection of critical values, sample size allocation and re-estimation, and selection of an interim sample size. In order to calculate power, type I error rate, and

expected sample size for a design, these factors must be pre-specified. In Chapter 3 of this dissertation, I discuss and evaluate procedures for planning the studies described in Chapter 2. The goal is to achieve sound study design strategies that control the type I error rate while best maintaining the power and sample size advantages of the IPIA designs.

In Chapter 4, the results necessary for the IPIA design for multiple degree of freedom tests in the GLUM framework are derived. These tests consist of more complex hypotheses such as multiple group comparisons. The key new result is the exact conditional joint distribution of the first and second stage test statistics. The new exact distributions may be used to solve for power, type I error rate, and expected sample size in these study designs. An example will demonstrate the characteristics of the IPIA design in this more complex setting.

1.2 LITERATURE REVIEW

1.2.1 Introduction

Interim analysis (or interim monitoring) often takes place during the course of clinical trials to acquire knowledge to make decisions such as design modification or early stopping. Jennison and Turnbull (2000, Chapter 1) categorized reasons for conducting interim analyses into three loosely defined classes: *ethical*, *economic*, and *administrative*. One ethical consideration includes ensuring that patients are not exposed to unsafe, inferior, or ineffective treatments. Another ethical consideration is the need to reallocate resources to other promising treatments when a current study is unlikely to show a benefit. Economic reasons for conducting interim analyses also exist from the ability to stop a trial early. If a trial stops early with a positive result, a treatment can reach the public more quickly, saving time and resources as well as generating an expedited revenue source for sponsors. Conversely, an early stopping event can also be triggered by an ineffective treatment or faulty study design. In this case, resources may be saved by stopping a study unlikely to result in a positive outcome under reasonable conditions if carried to completion.

Administrative reasons for conducting interim analyses include determining if the experiment is following both the designed protocol and planned assumptions. Assumptions made during sample size planning often include values for outcome variability in quantitative data or incidence rate values for binary data.

Three main types of interim analysis are group sequential methods, stochastic curtailment, and sample size re-estimation. Group sequential methods are analyses in which groups of subjects are enrolled and analyzed sequentially. They are designed to shorten the expected length of a study by allowing early hypothesis decisions to be reached if true effect sizes are larger or smaller than anticipated. Stochastic curtailment is another method of shortening a study based on calculating probabilities of achieving hypothesis decisions conditional on accumulated observed data. Sample size re-estimation procedures cover a wide range of possible study designs. All include possible adjustment to the planned sample size of a study (increases and/or decreases) in light of new information concerning aspects of the study.

Specific to sample size re-estimation, methods vary based on what information may be used, when the information is used, and the decisions made as a result. Flexible designs allow the most freedom with few restrictions if the type I error rate is controlled. Adaptive designs restrict the study to pre-planned design modifications based only on information internal to the study. Internal pilot designs allow for modification of a study based only on re-estimation of nuisance parameters, such as the error variance for Gaussian outcomes. The sample size re-estimation methods also vary based on if rules are included for early stopping at interim stages.

1.2.2 Group Sequential Methods

Group sequential methods (GSMs) allow for interim analyses in ongoing studies with the possibility of early stopping with sequentially enrolled subject groups. Although these methods may be applied to any study of sufficient duration that is completed in stages,

research focus has been the use of GSMs in clinical trials. An important reason for this focus is to stop randomization of patients to a potentially inferior treatment when a significant treatment difference can be proven with high probability.

An early influence in sequential analysis was Wald's (1947) sequential probability ratio test (SPRT). The SPRT tests between two simple hypotheses by sampling observations while the likelihood ratio remains in an interval (a, b) for constants a, b chosen to approximately control type I and type II error rates. Armitage (1975) developed methods for fully sequential analysis in medical studies. In these methods, data must be enrolled in matched pairs and accumulating data monitored continuously. However, this was not proven to conform well and did not achieve widespread use. GSMs worked better within clinical trials settings and became a popular alternative with the release of papers by Pocock (1977) and by O'Brien and Fleming (1979). For normally distributed data with known variance, these papers presented clear approaches for two-sided group sequential testing that controls the type I error rate while maintaining power. In a basic group sequential design for a comparison of two treatments, a maximum number of stages (k), group-size (m), and critical values for each stage (c_i for $i \in \{1, \dots, k\}$) are pre-determined. Subjects are randomized to treatment with the constraint that for each stage, m subjects are assigned to each treatment. For stage i a standardized statistic, Z_i , is computed using data from the first i groups and the study stops with rejection of the null hypothesis, H_0 , if $|Z_i| \geq c_i$, and continues otherwise. At stage k , H_0 is accepted if $|Z_k| < c_k$.

Critical values are chosen to preserve the overall type I error rate (α), i.e., $\Pr\left\{|Z_i| \geq c_i \text{ for } i \in \{1, \dots, k\} \mid H_0\right\} = \alpha$. To preserve type I error rate, Pocock uses an adjusted, constant nominal significance level for each test and O'Brien and Fleming describe a procedure for the nominal significance level to increase as the study progresses. Other techniques exist varying in degrees of complexities and efficacy. One very simple method is the Haybittle-Peto test (Haybittle, 1971; Peto et al., 1976), which stops a trial at stage i

($i < k$) if $|Z_i| \geq 3$ and then uses the ordinary stopping bound at the final stage. The final stage could also be slightly modified to accurately control type I error rate. This method has gained traction when trial planners need a simple rule to stop a study only when a clear and strong effect is observed while paying little penalty in the final critical value. Due to multiple testing possibilities, the maximum sample size of $2km$ is determined by a procedure-specific inflation factor multiplied by the sample size from a corresponding fixed sample test.

Wang and Tsatis (1987) described a family of two-sided test designs, indexed by a parameter Δ ($0 \leq \Delta \leq 0.5$), for use in the general GSM framework. The family generalizes the Pocock and O'Brien-Fleming methods with $\Delta = 0$ giving the O'Brien-Fleming test and $\Delta = 0.5$ giving the Pocock test. The adjusted critical values depend on k , α , and Δ ; the maximum sample size inflation factors depend on k , α , β , and Δ where β represents the target type II error rate. Lan and DeMets (1983) introduced a flexible way to construct boundaries in group sequential methods using an α -spending function. The idea is to define a monotonely increasing function for the information fraction t ($0 \leq t \leq 1$): $\alpha(t)$ with $\alpha(0) = 0$ and $\alpha(1) = \alpha$, the desired type I error rate. This function characterizes the rate at which the error level α is spent. This method can approximately emulate the Pocock and O'Brien-Fleming boundaries, but allows for other methods, for variable timing, and number of analyses. A number of possible error spending functions have been proposed in the literature (Lan and DeMets, 1983; Hwang et al., 1990; Kim and DeMets, 1987; Jennison and Turnbull, 1989).

While classical group sequential designs allow for reductions in sample size by stopping early for large effect sizes, they offer no reduction in sample size (in fact a small increase occurs) under the null hypothesis (e.g., $H_0 : \theta = 0$). Stopping a study early for futility, while not as ethically important as stopping when a significant difference is proven, can be important for financial considerations and resource allocation when the chance of a

significant study is low. Gould (1983) proposed methods for early stopping only to accept the null if a test has a p -value greater than a fixed critical value. Pampallona and Tsiatis (1994) described a one parameter (Δ) class of boundaries for group sequential methods based on the family introduced by Wang and Tsiatis (1987) that can be used for any type I and type II error rate choices. Whitehead and Stratton (1983) and Whitehead (1997) described an alternate method known as the *triangular test* based on combining two one-sided tests. Jennison and Turnbull (2000, Chapter 5) compared these methods as well as providing some tables of constants. Since allowing the study to stop to accept H_0 for small effect sizes may have significant savings in time and cost, Jennison and Turnbull (2000, Chapter 5) recommend that the stopping bounds be considered in all group sequential two-sided tests. Lachin (2005) explored the use of futility monitoring plans based on conditional power within group sequential testing. The method has a single futility analysis at a specified information fraction (such as $T = 0.5$) amidst the other interim tests before, at, and after the futility analysis. Using O'Brien-Fleming bounds, the plan approximately controls the type I error rate and maintains power while adding sample size benefits under the null.

Spurrier (1982) presented two-stage tests of hypothesis in the general linear univariate model with normally distributed and independent errors, a special case of group sequential methods. He proposes an ad hoc sample size selection method with each stage being of size $0.6*n_0$ where n_0 is the sample size of a fixed sample test. In the method, the first sample leads only to decisions to stop for efficacy (if $F_1 \geq c_u$, reject H_0) or futility (if $F_1 \leq c_l$, accept H_0), or to take the final sample (if $c_l < F_1 < c_u$). The null hypothesis is then either accepted or rejected if the final test statistic, F_+ , is below or above critical value c_+ , respectively. Proposed strategies for selecting critical values c_l , c_u , and c_+ are given with a primary concerns of controlling the type I error rate and secondary concerns of maximizing power and minimizing expected sample size. Hewett and Spurrier (1983) described with detail two-stage tests in a variety of settings. They promoted two-stage tests as a

compromise between fixed sample and sequential methods with more stages, offering well defined theory with a reduction in expected sample size while minimizing uncertainty of sample size and duration of study.

While most group sequential methods rely on large sample critical values or known variances, some alternative critical value selection methods have been proposed and reviewed. One simple approach suggested by Pocock (1977) shown to work quite well when variance is unknown is to take the significance level of Gaussian derived critical values and use them along with sample size to calculate corresponding t distributed critical values. Since the t distribution takes into account the sample size used for estimation in the form of degrees of freedom, it better relates to the uncertainty of the variance estimate used in the test statistic. Although the statistics are sequences and hence have a joint relationship, this simple method approximately controls the type I error rate for group sequential designs. Additionally, Shao and Feng (2007) described a Monte Carlo method for calculation of critical values in a small sample group sequential studies. Through simulation they showed that their method works well at controlling the type I error rate and maintaining power with an expected increase in expected sample size.

Group sequential methods have also been further generalized in various ways. Jennison and Turnbull (1991, 1997) described distributional theory for group sequential t , χ^2 , and F tests. Methods have also been described to allow for flexibility in the number, timings, and sizes of looks (Jennison and Turnbull, 2001).

1.2.3 Stochastic Curtailment

Study curtailment describes the idea that an experiment can be stopped once the outcome is inevitable; that is, further data collected within the study can not affect the final decision. In certain studies such as ones with normally distributed outcomes, there cannot be absolute certainty in an outcome as long as more sample size can be taken. A study may be

stopped, however, if the outcome while not inevitable, is highly probable. This can increase the efficiency of a study by decreasing expected sample size.

One approach to stochastic curtailment is the conditional power (CP) approach. CP is defined here as the probability of a statistically significant result (rejecting H_0) at the end of a study given a true value of the effect size and conditional on data already observed. Let $P_i(\theta)$ be the CP at some stage i for effect θ . A method described by Lan et al. (1982) defines a formal stopping rule where H_0 is rejected if $P_i(\theta_0) \geq c$, for a constant c such as 0.8 or 0.9. The logic behind this method is that the test will not likely accept the null at this point even if it is true. Alternatively, a test could be stopped for futility (accept H_0) if $1 - P_i(\theta_1) \geq c'$ where θ_1 is an alternative of interest. Proschan et al. (2006, Chapter 3) noted that, under a futility scenario, an estimate of nuisance parameters such as the sample variance could be used to recalculate *unconditional* power of the study. A low value implies an uninformative acceptance of H_0 and is further evidence to curtail the study. Under the scenario of a low CP and a high unconditional power, continuing the study may be useful to clearly differentiate between the hypotheses. A criticism of the CP stopping methods (Jennison and Turnbull, 2000, Chapter 10; Dmitrienko and Wang, 2006) is that they are based on calculations under only specific values of θ and ignore information about the effect size from current data. For example, an overly optimistic value of θ_1 would make a study difficult to stop for futility despite unpromising results.

Another form of stochastic curtailment known as the predictive power (PP) method utilizes a mixture of Bayesian and frequentist ideas. Jennison and Turnbull (2000, Chapter 10) described the approach, which averages conditional power over values of effect θ with weighting corresponding to current belief: a posterior distribution given the prior distribution and the observed data. This method gives an informative probability of success or failure in a study and, like the CP method, formal rules can be developed for early stopping for efficacy or futility. The method is described by Choi et al. (1985) and Spiegelhalter et al.

(1986) for binary endpoints. Choi and Pepple (1989) applied the Bayesian-frequentist approach to normally distributed data. Jennison and Turnbull (2000, Chapter 10) and Bernardo and Ibrahim (2000) also discussed the mixture approach in general settings. Criticisms of the method include the lack of a clear frequentist interpretation and that it is inconsistent with Bayesian principles (Jennison and Turnbull, 1990, Chapter 10).

A third method, described by Dmitrienko and Wang (2006), introduces a family of Bayesian stopping bounds by extending a *Bayesian predictive* method proposed by Geisser (1992). The paper reviews and compares the methods for stochastic curtailment. Dmitrienko et al. shows that the Bayesian and Bayesian-frequentist methods typically allow higher probability of early stopping with the pure Bayesian method being more sensitive to the choice of prior distribution. Dmitrienko et al. (2005, Chapter 4) provided SAS macros for the computation of stochastic curtailment stopping bounds for the three methods.

1.2.4 Sample Size Re-Estimation

1.2.4.1 Introduction

Sample size re-estimation (SSR) procedures differ from traditional and classical group sequential methods. This difference occurs as at least some of the information accrued during a study (possibly external to the study) is used to determine the size of future sampling. Historically, SSR proposals go back over 60 years. Some of the first research was proposed by Stein (1945), who introduced a two-stage procedure for normally distributed data where sample size is computed based on variance information contained in a first stage of the analysis. Later, Anscombe (1953), described a fully sequential procedure with sample size re-calculated repeatedly based on updated variance estimates. More recently, increased interest in the topic for application in clinical trials has led to many variations on the theme to make the design of a clinical trial more flexible and/or adaptive. Uncertainties about factors affecting power such as patient variation, treatment effect size, recruitment rates, or

event rates have led to researchers desiring the ability to make midcourse adjustments to the sample size of the study.

SSR methods vary by the number of stages used, the allowance for early stopping for efficacy and/or futility, the information used for re-estimation, if the adaptation protocol must be pre-specified, and if sample sizes are allowed to decrease. The great volume of recent research and sometimes lack of clear definitions and delineations has led to a confusion in terminology for similar methods. Different types of SSR methods will be introduced in this section with a focus of clarifying the similarities and differences that exist between them.

1.2.4.2 Flexible Designs

Flexible designs are study designs that permit mid-trial modifications with very little restriction. Information for study modifications can come from information internal or external to the trial. Also, adaptation does not need pre-specification. However, a major design consideration for flexible designs is to maintain the type I error rate in order to better maintain study validity. Flexible designs specifically will not be covered here; instead a meaningful subset will be discussed: *adaptive* designs.

1.2.4.3 Adaptive Designs

Recently, there has been great interest in the development of adaptive design (AD) methodology. ADs for clinical trials offer researchers flexibility to redesign trial procedures and analysis at interim stages. Current research, however, has created a confusion in terminology as many types of study modification are referred to as *adaptive*. In Spring 2005, a PhRMA working group on ADs in clinical drug development was formed to investigate and facilitate the acceptance and usage of these design methods. An Executive Summary of the group's findings (Gallo et al., 2006) defined an *adaptive design* as "a clinical study design that uses accumulating data to decide how to modify aspects of the study as it continues, without undermining the validity and integrity of the trial." This definition will be used

when referring to ADs here. Under this definition, the adaptations only use information from accumulating data internal to the trial as opposed to flexible designs which can also incorporate external information. The PhRMA working group also stresses that the changes should be made "by design" and not undertaken on an ad hoc basis. The definition makes it clear that adaptive designs are not meant to be a remedy for poor planning. Rather, ADs are meant to be designed study enhancements aimed at maintaining study validity and integrity while increasing efficiency of drug development and utilization of resources.

Bauer and Köhne (1994) and Proschan and Hunsberger (1995) were two of the early papers describing AD methods for adapting studies while maintaining type I error rate controls. Bauer and Köhne used a weighted Fisher's combination test for a two-stage one-sided test with possible early stopping and SSR based on effect size. Alternatively, Proschan and Hunsberger based their test on a conditional error approach: overall type I error rate is controlled as long as a second stage test conditional on the first stage results maintains the type I error rate. Wassmer (1998) showed that for two stages and one sided hypotheses, the methods of Bauer and Köhne (1994) and Proschan and Hunsberger (1995) are extremely similar in power and expected sample size. Other related adaptive methods include the methods described by Lehman and Wassmer (1999) and Cui et al. (1999). Both approaches use classical group sequential stopping boundaries with updating of sample size based on data observed in a first study and use fixed, predetermined weights to combine stage-wise results. The Lehman and Wassmer approach combines p -values using the inverse normal approach with fixed, predetermined weights (usually equal across stages). All of the methods above assume the variance is known for the study.

A number of issues have been raised concerning the use of adaptive designs. An obvious concern is the use of the observed treatment effect to re-estimate needed sample size. This issue will be discussed in section 1.2.4.4. Another issue is the potential abuse of weighting schemes in extreme samples that could potentially result in a significant test result

of a positive effect for a negative estimate (Proschan and Hunsberger, 1995; Burman and Sonesson, 2006). The weighting schemes used to protect type I error rate also violate basic sufficiency principles since observations from different stages are given different weightings (Jennison and Turnbull, 2003). Tsiatis and Mehta (2003) and Jennison and Turnbull (2006) argued that while adaptive designs have a place for preserving a study if unplanned analyses are conducted, group sequential methods offer more efficiency under reasonable conditions.

1.2.4.4 Sample Size Re-Estimation Based on Conditional Power

Conditional power (CP) has been proposed as a tool for the recalculation of sample size in clinical trials for adjusting study power (Proschan and Hunsberger, 1995). Two ways exist using interim data to calculate a probability of rejection in the trial given the results observed thus far. The first type of probability calculation assumes the true effect size, θ , is equal to the value the study was originally powered to detect, θ_1 . The other type assumes the true effect size is the observed estimate, $\hat{\theta}_i$, for an interim stage i .

The logic supporting the SSR is the adjustment of the sample size to that needed to maintain study power at the target rate. Often, this kind of calculation includes the revised estimates of nuisance parameters for SSR purposes (Denne, 2001). In this sense, the SSR is similar to an internal pilot technique. The difference lies in that regardless if θ_1 or $\hat{\theta}_i$ is assumed to be the true effect in the calculations, the calculations depend on the observed value of the test statistic at the interim stage. While different conceptually, both kinds of CP calculations raise questions in this context.

The less controversial use would be to adjust the sample size of the study to maintain target power at $\theta = \theta_1$ (Denne, 2001). For this method, due to conditioning on the interim test statistic, the study is resized to allow for room to 'catch up' if observed effects are lower than desired. On the other hand, if the effects are higher than anticipated, this practice can save resources if the planned sample size is decreased in order to maintain target power.

However, many researchers and even sponsors might prefer to keep the positive interim stage data and increase their overall probability of success in the trial.

If instead the study is re-powered to achieve target power at $\theta = \hat{\theta}_i$ (Cui et al., 1999), more issues are raised. In this case, the study is repowered at a new hypothesis of $\theta = \hat{\theta}_i$. If the sample size is increased, then statistically, the effect of interest is decreased. Uses for the procedure such as flexibility to external factors and cases where an effect of interest is unclear have been described by various researchers. In the case where sample size is decreased due to interim analysis, the test could be underpowered to detect the originally planned effect size if it is in fact true.

1.2.4.5 Internal Pilot Designs

A poor variance value used in sample size calculations can greatly impact the power of a clinical trial. A value that is too low leads to an underpowered study with a small chance of success regardless of the treatment's efficacy. Alternatively, a value that is too large leads to a waste of money and other resources in an overpowered trial. Stein (1945) introduced a two-stage t -test design with power independent from the variance. This technique updates the sample size at an interim stage using only the observed sample variance. The final test statistic uses information from all subjects for treatment effect, but only the variance estimate from the first sample. For a two group comparison, the final test statistic under the null hypothesis follows a t distribution with $n_1 - 2$ degrees of freedom where n_1 is the total sample size at the interim stage. A criticism of the method is that it throws away information about the variance from the second sample. Proschan and Wittes (2000) noted that the technique is not robust to possible changes in variance during the course of the trial. Also, Coffey and Muller (1999) showed Stein's method does not perform well when the second sample is large compared to the first.

Building on Stein's two-stage test, Wittes and Brittain (1990) introduced internal pilot (IP) designs for two groups with normally distributed outcomes. The researchers modified

Stein's method by using the pooled variance from all subjects in the final test statistic and treating the final sample size as if it were fixed. Simulation was used to show power and type I error rate for this test for an example using a preplanned sample size of 86 and an internal pilot using half of the preplanned sample. For large samples and an upward-restricted sample size adjustment design, they concluded that the type I error rate and power were well preserved.

Birkett and Day (1994) explored the use of different sizes for the interim stage rather than half of the initial fixed sample estimate. This design also allowed for decreases in the final sample size. The conclusion was reached that as long as there are enough degrees of freedom (~ 20) in the IP stage, the type I error rate and power are close to target levels. Coffey and Muller (1999) showed by counterexample that significant type I error rate inflation (up to 14% in their example) can still occur in test studies under this scenario. Coffey and Muller determined that the choice of internal pilot size and other design parameters can strongly affect results and should be inspected during planning for specific studies. Despite the potential benefits in power properties or sample size savings, the risk of type I error rate inflation, caused by a downward biased variance estimate (Proschan and Wittes, 2000; Miller, 2005) offsets the benefits in the minds of many researchers (Kieser and Friede, 2000) and regulatory agencies (ICH Topic E9 Guideline, Section 4.4). This risk has led many researchers to propose methods to control the type I error rate. These methods can be separated by if blinding is maintained on treatment allocation at the interim analysis, and by if the test statistic or critical value is modified to preserve the type I error rate.

From a regulatory standpoint, methods that keep the treatment group allocation blinded may be preferred to those that require unblinding (ICH Topic E9 Guideline, Section 4.4). For blinded sample size re-estimation, Gould and Shih (1992) and Zucker et al. (1999) suggested using the one-sample variance estimator with a simple adjustment based on the planned treatment effect of interest. Kieser and Friede (2003) showed that this approach

approximately controls the type I error rate when the true treatment difference is close to the prespecified difference. A disadvantage of the blinded methods is that the one-sample variance changes in relation to the treatment effect, which could cause inflation of sample sizes if the treatment effect is larger than thought. Friede and Kieser (2001) note that the sample size inflation is small when the true effect is close to the prespecified effect.

Many methods to control the type I error rate with unblinding have been proposed. Miller (2005) points out that the decision on whether or not to use blinded procedures should be made on a case-by-case basis and notes that careful control of information and the use of an independent statistician can mitigate potential biases. Stein's method controls the type I error rate by only using information from the first sample for the variance estimate. Zucker et al. (1999) proposed an alternate method where only the information about the variance independent from the IP stage is used in the final test statistic. This method controls the type I error rate both conditionally and unconditionally. Denne and Jennison (1999) proposed a method based on Stein's test that uses all information about the variance, but includes a degree of freedom adjustment to the final test statistic that does not guarantee bounding of type I error rate, but appears to work well in general. Proschan and Wittes (2000) introduced a method that uses an unbiased variance estimate by fixing weights between the IP stage and the second stage portions of the final variance estimate. Coffey and Muller (2001) introduced a *bounding method* which alters the critical value so that the maximum type I error rate inflation is equal to the target rate. The method by Miller (2005) adjusts the normal variance estimate to control the type I error rate.

Recent review papers by Proschan (2005) and Friede and Kieser (2006) review internal pilot designs for continuous and dichotomous outcomes. In the continuous case in which most research addresses the independent groups *t*-test setting. In order to accommodate more complex designs, Coffey and Muller (1999, 2000, 2001) have extended the idea of internal pilots into any univariate linear model with fixed predictors and Gaussian errors.

The researchers also derived computable forms for the exact distribution for the test statistic, which includes *t*-tests as special cases. Kairalla et al. (2007) released a free software package based in SAS/IML[®] (SAS Institute, 2004) for exact power, type I error rate, and expected sample size calculations for a wide range of internal pilot designs. For binary outcomes, Proschan (2005) describes the possibility of an underpowered study if the control event rate is overestimated. The paper describes two methods for re-estimating the sample size, both with asymptotic validity. In an unblinded method, the control event rate can be re-estimated and used for SSR. Alternatively, Gould (1992) described a blinded SSR procedure for binary data based on the overall event proportion. Internal pilot methods have also been extended into other settings of interest including ordinal data (Bolland et al., 1998), time-to-event data (Whitehead et al, 2001), and repeated measures (Shih and Gould, 1995; Lake et al., 2002; Zucker and Denne, 2002; Coffey and Muller, 2003).

1.2.4.6 Review of Related Topics

For clinical trials with Gaussian outcomes, IP designs allow for an update to sample size based on an error variance estimate taken at an internal stage. While studies may be lengthened or shortened by this estimate, the main objective is to ensure that the study is sufficiently powered to detect an effect size of interest. Group sequential methods, on the other hand, are designed to allow for a reduction in sample size if effect sizes deviate substantially from anticipated sizes. A successful combination of GSMs with IP based sample size re-estimation would allow for early stopping due to effect size differences and also help assure correctly powered studies for an effect of interest with respect to the true variance, a nuisance parameter. There have been a number of papers considering procedures for combining GSM and IP studies to obtain their respective benefits.

Stein's (1945) two-stage design was a strong early influence to SSR in sequential procedures. Baker (1950) and Hall (1962) introduced similar sequential tests based on the sequential probability ratio test (SPRT; Wald, 1947) incorporating information about the

variance using a single sample estimate. Arghami and Billard (1992) defined a partial sequential procedure also based on the SPRT and a Stein-like variance estimate. Hochberg and Marcus (1983) described a three-stage test for a one-sided, two-group comparison. This comparison uses variance information from a first sample to determine sample sizes for two testing stages. All of these procedures share the disadvantage of only incorporating early stage variance information into the test statistics.

Facey (1992) described a Phase 2 trial design using the triangular test stopping bounds. She compared the use of powering to absolute or standardized treatment differences. Type I error rate inflation was high for the absolute differences and more reasonable for the standardized differences in the cases considered (max type I error rate of 0.059 for target 0.05). Gould and Shih (1998) used a blinded variance estimate from the initial stage to fix future sample sizes. The procedure only allows for sample size increases to the group sequential procedure if the variance estimate is at least a constant factor larger than the planning value (increase sample size if $\hat{\sigma}_1^2 \geq \lambda \sigma_0^2$ with $\lambda = 1.33$, for example). A few methods are explored, such as redistributing the sample sizes to match the originally planned information times, or allowing the sample sizes to vary in pre-planned or unplanned manners. They concluded through simulation, with a small fraction of error dedicated to the first testing stage, that the procedure works adequately with two testing stages. Whitehead et al. (2001) explored through simulation a method similar to Gould (1998) for comparing effects from two groups by updating estimates of the standardized difference, δ_1/σ^2 , where δ_1 is the effect of interest and σ^2 the common variance. The study is first planned to detect $\theta_1 = \delta_1/\sigma_0^2$, which can then be revised by repowering to detect $\theta_1 = \delta_1/\hat{\sigma}_1^2$ using an estimate of σ^2 from an interim stage. The paper asserts that decision making will be flexible and up to a Steering Committee, but for simulation purposes created one possible strict study protocol. The use of both unblinded and blinded variance estimators were examined and similar results were concluded. The results were generally of a large sample nature (smallest average

sample scenario was $n = 92$). Despite the large samples, type I error rate inflation occurred in simulations they ran with or without SSR (up to 0.032 for target of 0.025). The authors noted that a large problem is that asymptotic results underlying sequential theory only become accurate for very large samples.

Denne and Jennison (2000) proposed a group sequential t -test with sample size update. This was based on the variance for a two-sided single group test of mean with early stopping to reject the null. A test based only on a Stein-like first variance estimate was first described. This method was used as a stepping stone to define a test procedure where the maximum sample size is recalculated at each stage with updated variance estimates. The remaining sample is then split based on the number pre-planned number of testing stages. Testing is not done at the first stage if the originally planned first stage testing fraction is not met. Thus, a two testing stage procedure could have three or more stages in total. In the calculations for critical values and sample size adjustments, both a type I error rate spending function and a degree of freedom correction are used to reflect the uncertainty of the variance estimates. The "effective" number of degrees of freedom at stage i is defined to be $n_1 + \epsilon(n_i - n_1) - 1$ for $0 \leq \epsilon \leq 1$ and n_1 the first stage sample size. Based on calculations for several examples, $\epsilon = 1/4$ is recommended to approximately achieve target error rates. For tests with two and five stages, Denne and Jennison showed by a combination of simulation and numerical integration that the procedure works reasonably well, especially when n_1 is large (for example, ≥ 20). For the two-stage test with low first stage sample ($n_1 = 5$) type I error rate inflation in this example can occur with a worst case considered of 0.062 for a target rate of 0.05. Morgan (2003) considered sample size re-estimation in group sequential trials with the goal of extending the idea for use in group-sequential response-adaptive designs for Gaussian data. Morgan compared the performance of techniques similar to the ones described by Denne and Jennison (2000) to conclude through simulation that a

design using updated variance estimates at each stage has better power and sample size properties.

Another approach to clinical trial monitoring with nuisance parameter based sample size adjustment is the information based approach described by Mehta and Tsiatis (2001) or Tsiatis (2006). Since statistical precision is determined by the amount of statistical information, a study should continue until the needed statistical information level is reached. At this point the study will closely achieve the desired statistical power. Mehta and Tsiatis described the method for use within group sequential designs that allow for early stopping while updating the estimated maximum sample size at each analysis stage as nuisance parameter estimates are updated. Group sequential stopping bounds along with an inflation factor on needed information (and hence needed sample size) due to multiple testing were advocated. They used standardized test statistics with critical boundary determination based on the error spending technique used. Large samples are needed for this design in order to avoid type I error rate inflation caused by asymptotic properties in the distribution of the test statistics and boundary point calculations as well as from the of a downwardly biased variance estimate in study stages following the first. This is the same cause of type I error rate inflation found in unadjusted internal pilot studies; see Proschan and Wittes (2000) or Miller (2005) for details.

Most work has dealt only with the one or two group t -test scenario. This work represents an intersection with the topics of this dissertation. Many of the results promote general techniques for group sequential designs and are typically based on underlying large sample assumptions to account for designs adjustments made in the trials. Group sequential designs typically have a primary goal of reducing average sample size by frequently monitoring studies in order to stop if effect sizes are larger or smaller than planned. Internal pilots, typically only needing one interim analysis, attempt to check and correct for possible misspecification of nuisance parameters in order to secure power levels for a study. Possible

sample size reductions are a secondary benefit. The primary focus of this dissertation is maintaining power by updating sample size needs, while incorporating the benefits of group sequential theory by allowing the possibility of early stopping at the interim stage.

1.3 SUMMARY

Many prospective research studies and even clinical trials are not large enough for asymptotic properties to hold. Researchers in small sample continuous outcome settings need the ability to control type I error rate and maintain power over possible values of the error variance, a nuisance parameter, while minimizing sample size needs. These small sample settings can be one or two group studies (t -tests), multiple group comparisons, or other designs. Distributional knowledge and effective protocols in these settings would be valuable to study designers.

Currently methods do not exist for exact theory calculations of power, type I error rate, and expected sample size in small samples for an internal pilot design with interim analysis for early stopping. These calculations could greatly increase the efficiency of study planning through fast direct calculation. Useful and accurate sample size re-estimation and critical value selection criteria that can control the type I error rate while maintaining power and minimizing expected sample size are also unclear in the small sample settings, with most methods being asymptotic results.

For the two-stage IPIA method, exact theory in the GLUM setting with Gaussian errors and fixed predictors are derived. The theory may be applied accurately in small sample settings and a wide range of study designs within the GLUM framework. Logically and computationally, this method collapses to the unadjusted IP design detailed in Coffey and Muller (1999) for early stopping regions set to the null space as well as to a two-stage group sequential test when sample size re-estimation not allowed.

CHAPTER 2. INTERNAL PILOT WITH INTERIM ANALYSIS FOR SINGLE DEGREE OF FREEDOM HYPOTHESIS TESTS

SUMMARY

In this chapter, I introduce the proposed model of an internal pilot with interim analysis (IPIA) design, discuss sample size re-estimation technique, and derive the exact distributional theory needed for planning studies with single degree of freedom tests. The exact distributional theory allows computation of power, type I error rate, and expected sample size for one and two group comparisons with unknown, common variances and other single degree of freedom hypothesis univariate linear model study designs with fixed predictors and Gaussian errors. Examples compare study characteristics with a fixed sample design as well as with the internal pilot and two-stage group sequential designs, all of which can be seen as special cases within the IPIA framework.

2.1 INTRODUCTION

2.1.1 Motivation

When planning clinical trials and other studies, researchers would like to ensure they have an appropriate sample size to detect an effect of interest for a given target type I error rate and power. Researchers and sponsors would also like to have the ability to reach early decisions when hypothesis outcomes are clear. Often times studies consist of one or two group effect size comparisons. Much of the current results promote general techniques for group sequential type designs and are typically based on underlying large sample assumptions to account for design adjustments made during the trials.

Group sequential designs have a primary goal of reducing average sample size by frequently monitoring studies in order to stop early if effect sizes are larger or smaller than planned. Internal pilots, typically only needing one interim power analysis, have the alternative goal of checking and correcting for possible misspecification of nuisance parameters in order to secure power levels for a study. Possible sample size reductions are a secondary benefit. By developing procedures and theory for two-stage designs with interim analyses, I focus primarily on the goal of maintaining power by updating sample size needs, while also incorporating the benefits early stopping procedures. Exact distributional results with computable formulae for power and sample size would allow researchers to accurately explore properties for such designs, even in small samples, before undertaking a study. The exact theory would allow for efficient study planning without the need for simulations, even in small sample studies.

The importance of small sample theory is explicitly highlighted within the NIH Roadmap (Clinical and Translational Science Awards, RFA-RM-07-002 U54). Also, while large sample clinical trials get a lot of attention, they are often based on numerous small sample studies. The results of this chapter can be used to examine exact properties for many study designs in a two-stage framework, including the information based approach. The use of the exact theory can facilitate studying and comparing properties of new methods in order to ascertain ones with the most desirable features for a particular study. For example, a new method for determining boundary points could lead to a more unbiased testing procedure in small sample studies. The use of the exact theory for procedure comparison will be explored in Chapter 3.

2.1.2 Literature Review

For a prospective study with Gaussian outcomes, an internal pilot (IP) design allows for an update to sample size based on an error variance estimate taken at an interim stage. While studies may be lengthened or shortened from their pre-planned size by this estimate, the main

objective of an internal pilot design is to ensure that the study is sufficiently powered to detect an effect size of interest. Group sequential (GS) methods, on the other hand, are designed to allow for a possible reduction in pre-planned sample size due to early stopping for efficacy or futility if effect sizes deviate substantially from anticipated magnitudes. Current research is looking at ways to simultaneously obtain the benefits of both approaches, i.e., combine the early stopping benefits of GS methods with sample size re-estimation methods (such as IPs) protecting against misspecification of nuisance parameters. There have been a number of papers considering procedures for combining GS and IP studies to simultaneously obtain their respective benefits.

Stein's (1945) two-stage design, which used variance information only from the first stage, was a strong early influence to sample size re-estimation in sequential procedures. Baker (1950) and Hall (1962) introduced similar sequential tests based on the sequential probability ratio test (SPRT; Wald, 1947) incorporating information about the variance using a single sample estimate. Arghami and Billard (1992) define a partial sequential procedure also based on the SPRT and a Stein-like variance estimate. Hochberg and Marcus (1983) describe a three-stage test for a one-sided, two-group comparison using variance information from a first sample to determine sample sizes for two testing stages. All of these procedure have in common the disadvantage of only incorporating early stage variance information into the test statistics.

Facey (1992) described a Phase 2 trial design using the triangular test stopping bounds (Whitehead and Stratton, 1983). She compared the use of powering to absolute or standardized treatment differences. Type I error rate inflation was high for the absolute differences and more reasonable for the standardized differences in the cases considered (maximum type I error rate of 0.059 for target 0.05). Gould and Shih (1998) used a blinded variance estimate from the initial stage to fix future sample sizes. The procedure only allows for sample size increases to the group sequential procedure if the variance estimate is at least

a constant factor larger than the planning value (increase sample size if $\hat{\sigma}_1^2 \geq \lambda \sigma_0^2$ with $\lambda = 1.33$, for example). In this case, they explore a few methods such as redistributing the sample sizes to match the originally planned information times, or allowing them to vary in pre-planned or unplanned manners. They concluded through simulation, with a small fraction of error dedicated to the first testing stage, that the procedure works adequately with two testing stages. Whitehead et al. (2001) explored through simulation a method similar to the one described by Gould and Shih (1998) for comparing effects from two groups by updating estimates of the standardized difference, δ_1/σ^2 , where δ_1 is the effect of interest and σ^2 the common variance. The study is first planned to detect $\theta_1 = \delta_1/\sigma_0^2$, which can then be revised by repowering to detect $\theta_1 = \delta_1/\hat{\sigma}_1^2$ using an estimate of σ^2 from an interim stage. In the paper, the authors assert that decision making will be generally flexible and up to a Steering Committee, but for simulation purposes they created a possible strict study protocol. They examined the use of both unblinded and blinded variance estimators and concluded similar results. The results were generally of a large sample nature (smallest average sample scenario was $n = 92$). Despite the larger samples, type I error rate inflation occurred in simulations they ran with or without sample size re-estimation (up to 0.032 for target of 0.025). The authors noted that asymptotic results underlying sequential theory only become accurate for very large samples.

Denne and Jennison (2000) proposed a group sequential t -test with sample size update based on the variance for a two-sided single group test of mean with early stopping to reject the null. They first described a test based only on a Stein-like first variance estimate. They used this method as a stepping stone to define a test procedure where the maximum sample size is recalculated at each stage with updated variance estimates and remaining sample split based on the pre-planned number of testing stages. For the procedure, testing is not done at the first stage if the originally planned first stage testing fraction is not met. Thus, a two testing stage procedure could have three (or more) stages in total. In the calculations for

critical values and sample size adjustments, they used a type I error rate spending function and an ad hoc degree of freedom correction to reflect the uncertainty of the variance estimates used. The "effective" number of degrees of freedom at stage i is defined as $n_1 + \epsilon(n_i - n_1) - 1$ for $0 \leq \epsilon \leq 1$ and n_1 the first stage sample size. Based on calculations for several examples, $\epsilon = 1/4$ is recommended to approximately achieve target error rates. For tests with two and five stages, Denne and Jennison showed by a combination of simulation and numerical integration that the procedure works reasonably well, especially when n_1 is large (say ≥ 20). For the two-stage test with low first stage sample ($n_1 = 5$) type I error rate inflation in their example can occur with a worst case considered of 0.062 for a target rate of 0.05.

Morgan (2003) considered sample size re-estimation in group sequential trials with the goal of extending the idea for use in group-sequential response-adaptive designs for Gaussian data. Morgan compared the performance of similar techniques to those described by Denne and Jennison (2000) and concluded through simulation that the use of updated variance estimates at each stage had beneficial power and sample size properties.

Another approach to clinical trial monitoring with nuisance parameter based sample size adjustment is the information based approach described by Mehta and Tsiatis (2001) or Tsiatis (2006). Since statistical precision is determined by the amount of statistical information, a study should continue until the needed statistical information level is reached. At this point the study will closely achieve the desired statistical power. Mehta and Tsiatis described the method for use within group sequential designs that allow for early stopping while updating the estimated maximum sample size at each analysis stage as nuisance parameter estimates are updated. Group sequential stopping bounds along with an inflation factor on needed information (and hence needed sample size) due to multiple testing were advocated. They used standardized test statistics with critical boundary determination based on the error spending technique used.

Large samples are needed for this design in order to avoid type I error rate inflation caused by asymptotic properties in the distribution of the test statistics and boundary point calculations. Another cause of type I error rate inflation in small samples for this design comes from the use of a downwardly biased variance estimate in study stages following the first. This is the same cause of type I error rate inflation found in unadjusted internal pilot studies; see Proschan and Wittes (2000) or Miller (2005) for details.

2.2 THE IPIA MODEL AND PROPERTIES

2.2.1 Notation

Notational conventions will be followed as described in Muller and Stewart (2006, Chapter 1). An $r \times 1$ vector (always a column) is written \mathbf{a} , and an $r \times c$ matrix is written $\mathbf{A} = \{a_{j,k}\}$, with transpose \mathbf{A}' . For full rank matrix \mathbf{A} , the inverse of the transpose equals the transpose of the inverse and I will use $\mathbf{A}^{-t} = (\mathbf{A}')^{-1} = (\mathbf{A}^{-1})'$. $\mathbf{1}_r$ always represents an $r \times 1$ vector of 1's and $\text{Dg}(\mathbf{x})$ represents a diagonal matrix with (j, j) element x_j . Furthermore, define \mathbf{I}_r as the $r \times r$ identity matrix with $\mathbf{I}_r = \text{Dg}(\mathbf{1}_r)$. The direct (Kronecker) product is defined as $\mathbf{A} \otimes \mathbf{B} = \{a_{j,k}\mathbf{B}\}$.

Detailed information about all random variables discussed in this paper can be found in Johnson et al. (1994, 1995). The vector $\mathbf{x} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ indicates that random vector \mathbf{x} ($n \times 1$) has a vector (multivariate) Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. For $\boldsymbol{\Sigma}$ less than full rank, \mathbf{x} has singular vector Gaussian distribution, written as $\mathbf{x} \sim S\mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Writing $X \sim \chi^2(\nu, \omega)$ indicates that X follows a non-central chi-square distribution, with ν degrees of freedom and noncentrality ω . Likewise, writing $X \sim F(\nu_1, \nu_2, \omega)$ indicates that X follows a noncentral F distribution with numerator degrees of freedom ν_1 , denominator degrees of freedom ν_2 , and noncentrality ω . Writing $\chi^2(\nu)$ or $F(\nu_1, \nu_2)$ implies $\omega = 0$. More generally, writing $X \sim \chi_T^2(\nu; t_L, t_U)$ indicates that X follows a doubly-truncated central chi-square distribution with ν degrees of freedom, truncated to the interval $[t_L, t_U]$ (Coffey and Muller, 2000). For random variable U with

parameters $\gamma_1 \dots \gamma_k$, indicate the cumulative distribution function (CDF) taken at u as $F_U(u; \gamma_1 \dots \gamma_k)$. As a special case, let $\Phi(z)$ indicate the CDF for the Gaussian(0,1) distribution, taken at z . Also $F_U^{-1}(\alpha; \gamma_1 \dots \gamma_k)$ indicates the α quantile of a random variable U with parameters $\gamma_1 \dots \gamma_k$.

2.2.2 The IPIA Model

The internal pilot with interim analysis (IPIA) models discussed in this paper can be viewed as generalizations of the two-stage internal pilot model in the GLUM framework as introduced in Coffey and Muller (1999), which includes the one and two sample t -tests as special cases. However, due to the possibility of early stopping, notational adaptations are necessary. In an IP design, N_+ ($n_{+,min} \leq N_+ \leq n_{+,max}$) is the random final sample size that is calculated using $\hat{\sigma}_1^2$, the variance estimate from the interim sample. For the IPIA model, N_+ ($n_1 \leq N_+ \leq n_{+,max}$) is also a random variable based on $\hat{\sigma}_1^2$ and fully determines the variable $N_2 = N_+ - n_1$. However, due to the possibility of early stopping, it is not necessarily the final sample size for the study. Let random variable N_w be the final sample size used for the study. Then $N_w = n_1 + N_2 \cdot \mathcal{I}(\text{continue})$ with \mathcal{I} an event indicator equal to 1 if a study is continued at the first stage. So

$$N_w = \begin{cases} n_1 & \text{if study stopped after first stage} \\ N_+ & \text{otherwise} \end{cases} . \quad (2.1)$$

The design leads to interest in different but intimately connected models. The combined model for the final analysis may be written as

$$\begin{matrix} \mathbf{y}_+ \\ N_+ \times 1 \end{matrix} = \begin{matrix} \mathbf{X}_+ \boldsymbol{\beta} \\ N_+ \times q \times 1 \end{matrix} + \begin{matrix} \mathbf{e}_+ \\ N_+ \times 1 \end{matrix}, \quad (2.2)$$

or

$$\begin{bmatrix} \mathbf{y}_1 \\ n_1 \times 1 \\ \mathbf{y}_2 \\ N_2 \times 1 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ n_1 \times q \\ \mathbf{X}_2 \\ N_2 \times q \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{e}_1 \\ n_1 \times 1 \\ \mathbf{e}_2 \\ N_2 \times 1 \end{bmatrix}, \quad (2.3)$$

with partitioning corresponding to the fixed n_1 and random N_2 observations in the first and second samples, respectively. The second sample of size $N_2 = N_+ - n_1$ shown above is only taken if study continuation is determined from the first sample. Also, the special case of $N_+ = n_1$ will cause the full model to collapse to the interim model. Model components include random observed \mathbf{y}_+ ($N_+ \times 1$) (independent sampling units as rows), design matrix of fixed form \mathbf{X}_+ , and unobserved \mathbf{e}_+ such that $\mathbf{e}_+ \sim \mathcal{N}_{N_+}(\mathbf{0}, \sigma^2 \mathbf{I}_{N_+})$. For computational convenience, random values of sample size, $N_+ = n_1 + N_2$, increase only in multiples of a replication factor, m . For example, a balanced 2-group study design would have $m = 2$. For some \mathbf{X}_0 ($m \times q$), assume $\mathbf{X}_1 = \mathbf{1}_{k_1} \otimes \mathbf{X}_0$ and $\mathbf{X}_2 = \mathbf{1}_{K_2} \otimes \mathbf{X}_0$, with fixed k_1 and random K_2 the number of replications in the first and second samples, respectively. Consequently, the columns of \mathbf{X}_1 and \mathbf{X}_2 span the same space (when $K_2 > 0$) and hence define $r = \text{rank}(\mathbf{X}_1) = \text{rank}(\mathbf{X}_2) = \text{rank}(\mathbf{X}_+)$. In order to simplify computations and some discussions, attention will usually be restricted to a full rank design, that is $\text{rank}(\mathbf{X}_0) = q$. The principles of linearly equivalent models allow the restriction without meaningful loss of generality.

The test of interest is $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, with \mathbf{C} a fixed $a \times q$ contrast matrix and $\boldsymbol{\theta} = \mathbf{C}\boldsymbol{\beta}$. Without loss of generality I assume $\boldsymbol{\theta}_0 = \mathbf{0}$ (see Lemma A.1). For a ‘scientifically important’ effect of interest ($\boldsymbol{\theta} = \boldsymbol{\theta}_1$), I seek a design that ensures a target type I error rate (α_t) with sample size appropriate to achieve target power (P_t).

Throughout, subscript $s \in \{1, +\}$ indicates a value for either the model based on the internal pilot (first) sample or the total combined sample (conditioned on $N_+ = n_+$). Error degrees of freedom are $\nu_s = n_s - r$. I use functional notation in many places to emphasize the dependence on n_s . For example, with the ‘hat’ matrix \mathbf{H}_s defined as

$$\mathbf{H}_s = \mathbf{X}_s(\mathbf{X}_s' \mathbf{X}_s)^{-1} \mathbf{X}_s' , \quad (2.4)$$

then,

$$\widehat{\boldsymbol{\theta}}(n_s) = \mathbf{C}(\mathbf{X}'_s \mathbf{X}_s)^{-1} \mathbf{X}'_s \mathbf{y}_s \quad (2.5)$$

and

$$\widehat{\sigma}^2(n_s) = \mathbf{y}'_s (\mathbf{I}_{n_s} - \mathbf{H}_s) \mathbf{y}_s / \nu_s \quad (2.6)$$

represent the unadjusted estimates of $\boldsymbol{\theta}$ and σ^2 for the model based on sample s . Similarly, define 'middle' matrix \mathbf{M}_s as

$$\mathbf{M}_s = \mathbf{C}(\mathbf{X}'_s \mathbf{X}_s)^{-1} \mathbf{C}' \quad (2.7)$$

and noncentrality δ_s as

$$\delta_s = \boldsymbol{\theta}' \mathbf{M}_s^{-1} \boldsymbol{\theta} \quad (2.8)$$

Then

$$\widehat{\delta}(n_s) = \widehat{\boldsymbol{\theta}}(n_s)' \mathbf{M}_s^{-1} \widehat{\boldsymbol{\theta}}(n_s) \quad (2.9)$$

is the observed hypothesis sum of squares for the model based on sample s . Hence, the unadjusted test statistics for the two stages are defined as

$$F(n_s) = \left[\widehat{\delta}(n_s) / a \right] / \widehat{\sigma}^2(n_s) \quad (2.10)$$

When there is no confusion, the functional aspects of the estimators will be implied with a subscript, e.g., $\widehat{\boldsymbol{\theta}}(n_s)$ is written $\widehat{\boldsymbol{\theta}}_s$ and $F(n_s)$ is written $F_1 = \left(\widehat{\delta}_1 / a \right) / \widehat{\sigma}_1^2$. I consider only testable hypotheses, which require full rank \mathbf{C} as well as $\mathbf{C}(\mathbf{X}'_s \mathbf{X}_s)^{-} (\mathbf{X}'_s \mathbf{X}_s) = \mathbf{C}$. Table 2.1 summarizes relevant dimensions, Table 2.2 summarizes parameters and constants, and

Table 2.3 summarizes design factors for the study.

Table 2.1: Dimensions

Symbol	Definition
n_1	Sample size, first stage
N_+	Total random sample size if study continued at first stage
N_2	Random second stage sample size, $N_+ - n_1$
N_w	Total random sample size used in study
q	Number of predictors and columns in \mathbf{X}_0
m	Replication factor, number of rows in \mathbf{X}_0
r	$\text{rank}(\mathbf{X}_0)$, $\text{rank}(\mathbf{X}_1)$, $\text{rank}(\mathbf{X}_+)$, and $\text{rank}(\mathbf{X}_2)$ when $N_2 > 0$
ν_s	Error df = $n_s - r$, $s \in \{1, +\}$
a	Number of rows in \mathbf{C} , hypothesis df for test statistics
k_1	Number of replications in first stage, n_1/m
K_2	Number of replications in second stage, random, N_2/m

Table 2.2: Parameters and constants

Symbol	Size	Definition and Properties
\mathbf{X}_0	$m \times q$	Fixed, known, base design matrix
\mathbf{X}_1	$n_1 \times q$	Fixed, known, first stage design matrix
\mathbf{X}_+	$N_+ \times q$	Final design matrix
\mathbf{X}_2	$N_2 \times q$	Stage 2 design matrix
$\boldsymbol{\beta}$	$q \times 1$	Primary parameters
\mathbf{C}	$a \times q$	Between-subject contrast matrix
$\boldsymbol{\theta} = \mathbf{C}\boldsymbol{\beta}$	$a \times 1$	Secondary parameters
$\boldsymbol{\theta}_0$	$a \times 1$	Null hypothesis values (can set to $\mathbf{0}$ WOLOG)
$\mathbf{M}_s = \mathbf{C}(\mathbf{X}'_s \mathbf{X}_s)^{-1} \mathbf{C}'$	$a \times a$	'Middle' matrix for stage s
$\mathbf{H}_1 = \mathbf{X}_1(\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1$	$n_1 \times n_1$	'Hat' matrix for first stage
$\mathbf{H}_+ = \mathbf{X}_+(\mathbf{X}'_+ \mathbf{X}_+)^{-1} \mathbf{X}'_+$	$N_+ \times N_+$	'Hat' matrix for second stage
σ^2	1×1	True variance
$\delta_s = \boldsymbol{\theta}' \mathbf{M}_s^{-1} \boldsymbol{\theta}$	1×1	Unscaled noncentrality for stage s
$\lambda_s = \delta_s / \sigma^2$	1×1	scaled noncentrality for stage s

Table 2.3: *Internal pilot with interim analysis notation*

	Symbol	Definition
Design Parameters	α_t	Target type I error rate
	P_t	Target power
	$\boldsymbol{\theta}_1$	'Scientifically Important' value of $\boldsymbol{\theta}$
	σ_0^2	Variance value used for planning
	n_0	Planned sample size for $\alpha_t, P_t, \boldsymbol{\theta}_1, \sigma_0^2$
	$f_l(n_+), f_u(n_+), f_+(n_+)$	Critical values determined by N_+
Sample Size	π	Proportion of n_0 used in internal pilot
Allocation	$n_1 = \pi n_0$	Internal pilot sample size
	$n_{+,max}$	Maximum size of final sample
Unknown Parameter	$\gamma = \sigma^2/\sigma_0^2$	Ratio of true to planning variances

2.2.3 IPIA Properties

This section presents model properties needed for future proofs and consideration.

Lemma 2.1 For the model in equation 2.3 interpreted as a fixed n_+ design, the following holds.

The following $n_+ \times n_+$ matrices are symmetric and idempotent, for any testable hypothesis, and have ranks of a , $(n_+ - r)$, $(n_1 - r)$ and n_2 :

$$\mathbf{A}_{h+} = \mathbf{X}_+(\mathbf{X}'_+\mathbf{X}_+)^{-1}\mathbf{C}'[\mathbf{C}(\mathbf{X}'_+\mathbf{X}_+)^{-1}\mathbf{C}']^{-1}\mathbf{C}(\mathbf{X}'_+\mathbf{X}_+)^{-1}\mathbf{X}'_+ \quad (2.11)$$

$$\mathbf{A}_{e+} = \mathbf{I}_{n_+} - \mathbf{X}_+(\mathbf{X}'_+\mathbf{X}_+)^{-1}\mathbf{X}'_+ \quad (2.12)$$

$$\mathbf{A}_{e1} = \begin{bmatrix} \mathbf{I}_{n_1} - \mathbf{X}_1(\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0}_{n_2 \times n_2} \end{bmatrix} \quad (2.13)$$

$$\mathbf{A}_{ep} = \mathbf{A}_{e+} - \mathbf{A}_{e1} . \quad (2.14)$$

Furthermore

$$\mathbf{A}_{h+}\mathbf{A}_{e+} = \mathbf{A}_{h+}\mathbf{A}_{e1} = \mathbf{A}_{h+}\mathbf{A}_{ep} = \mathbf{A}_{e1}\mathbf{A}_{ep} = \mathbf{0}. \quad (2.15)$$

See Coffey and Muller (1999) for details.

Extending the notation gives

$$\mathbf{A}_{h1} = \begin{bmatrix} \mathbf{X}_1(\mathbf{X}'_1\mathbf{X}_1)^{-t}\mathbf{C}'\left[\mathbf{C}(\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{C}'\right]^{-1}\mathbf{C}(\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}_{n_+ \times n_+} . \quad (2.16)$$

Or, equivalently, let $\mathbf{X}_{1*} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{0}_{n_2 \times q} \end{bmatrix} = \begin{bmatrix} \mathbf{1}_{k_1} \otimes \mathbf{X}_0 \\ \mathbf{0}_{n_2 \times q} \end{bmatrix}$ which gives

$$\mathbf{A}_{h1} = \mathbf{X}_{1*}(\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{C}'\left[\mathbf{C}(\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{C}'\right]^{-1}\mathbf{C}(\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_{1*} . \quad (2.17)$$

In turn, \mathbf{A}_{hs} , for $s \in \{1, +\}$, is idempotent, symmetric, and rank a (testable hypothesis).

Hence one can define \mathbf{V}_{hs} of dimension $n_+ \times a$ with

$$\mathbf{A}_{hs} = \mathbf{V}_{hs}\mathbf{V}'_{hs} \quad (2.18)$$

and

$$\mathbf{V}'_{hs}\mathbf{V}_{hs} = \mathbf{I}_a . \quad (2.19)$$

Also, since $\mathbf{A}_{ep} = \mathbf{A}_{e+} - \mathbf{A}_{e1}$ is idempotent, symmetric, and rank n_2 , one can define \mathbf{V}_{ep} of dimension $n_+ \times n_2$ with

$$\mathbf{A}_{ep} = \mathbf{V}_{ep}\mathbf{V}'_{ep} \quad (2.20)$$

and

$$\mathbf{V}'_{ep}\mathbf{V}_{ep} = \mathbf{I}_{n_2} . \quad (2.21)$$

Since it is symmetric and rank a , the matrix

$$\mathbf{M}_0 = \mathbf{C}(\mathbf{X}'_0\mathbf{X}_0)^{-1}\mathbf{C}' \quad (2.22)$$

can be written as $\mathbf{F}_0\mathbf{F}'_0$ for \mathbf{F}_0 of dimension $a \times a$ and rank a . This in turn implies that

$$\mathbf{M}_0^{-1} = \mathbf{F}_0^{-t}\mathbf{F}_0^{-1} . \quad (2.23)$$

For $k_s = n_s/m$ the number of replications at stage s , the fact that $\mathbf{M}_s = k_s^{-1}\mathbf{M}_0$ implies that

$$\mathbf{M}_s^{-1} = k_s\mathbf{F}_0^{-t}\mathbf{F}_0^{-1} . \quad (2.24)$$

Also $\mathbf{X}_s = \mathbf{1}_{k_s} \otimes \mathbf{X}_0$ implies that

$$\mathbf{X}'_s \mathbf{X}_s = k_s \mathbf{X}'_0 \mathbf{X}_0 . \quad (2.25)$$

From equation 2.18, $\mathbf{A}_{hs} = \mathbf{V}_{hs} \mathbf{V}'_{hs}$, with \mathbf{V}_{hs} of dimension $n_+ \times a$ and $\mathbf{V}'_{hs} \mathbf{V}_{hs} = \mathbf{I}_a$.

The matrices \mathbf{V}_{h+} and \mathbf{V}_{h1} can now be derived as follows:

$$\begin{aligned} \mathbf{A}_{h+} &= \mathbf{X}_+ (\mathbf{X}'_+ \mathbf{X}_+)^{-t} \mathbf{C}' \mathbf{M}_+^{-1} \mathbf{C} (\mathbf{X}'_+ \mathbf{X}_+)^{-1} \mathbf{X}'_+ \\ &= \mathbf{X}_+ (k_+ \mathbf{X}'_0 \mathbf{X}_0)^{-t} \mathbf{C}' \mathbf{M}_+^{-1} \mathbf{C} (k_+ \mathbf{X}'_0 \mathbf{X}_0)^{-1} \mathbf{X}'_+ \\ &= k_+^{-2} k_+ \mathbf{X}_+ (\mathbf{X}'_0 \mathbf{X}_0)^{-t} \mathbf{C}' \mathbf{F}_0^{-t} \mathbf{F}_0^{-1} \mathbf{C} (\mathbf{X}'_0 \mathbf{X}_0)^{-1} \mathbf{X}'_+ \\ &= k_+^{-1} \mathbf{X}_+ (\mathbf{X}'_0 \mathbf{X}_0)^{-1} \mathbf{C}' \mathbf{F}_0^{-t} \mathbf{F}_0^{-1} \mathbf{C} (\mathbf{X}'_0 \mathbf{X}_0)^{-1} \mathbf{X}'_+ \\ &= k_+^{-1} \left[\mathbf{X}_+ (\mathbf{X}'_0 \mathbf{X}_0)^{-1} \mathbf{C}' \mathbf{F}_0^{-t} \right] \left[\mathbf{F}_0^{-1} \mathbf{C} (\mathbf{X}'_0 \mathbf{X}_0)^{-1} \mathbf{X}'_+ \right] \end{aligned} \quad (2.26)$$

implies that

$$\mathbf{V}_{h+} = k_+^{-1/2} \mathbf{X}_+ (\mathbf{X}'_0 \mathbf{X}_0)^{-1} \mathbf{C}' \mathbf{F}_0^{-t} , \quad (2.27)$$

and

$$\begin{aligned} \mathbf{A}_{h1} &= \mathbf{X}_{1*} (\mathbf{X}'_1 \mathbf{X}_1)^{-t} \mathbf{C}' \left[\mathbf{C} (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{C}' \right]^{-1} \mathbf{C} (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_{1*} \\ &= k_1^{-1} \mathbf{X}_{1*} (\mathbf{X}'_0 \mathbf{X}_0)^{-1} \mathbf{C}' \mathbf{F}_0^{-t} \mathbf{F}_0^{-1} \mathbf{C} (\mathbf{X}'_0 \mathbf{X}_0)^{-1} \mathbf{X}'_{1*} \\ &= k_1^{-1} \left[\mathbf{X}_{1*} (\mathbf{X}'_0 \mathbf{X}_0)^{-1} \mathbf{C}' \mathbf{F}_0^{-t} \right] \left[\mathbf{F}_0^{-1} \mathbf{C} (\mathbf{X}'_0 \mathbf{X}_0)^{-1} \mathbf{X}'_{1*} \right] \end{aligned} \quad (2.28)$$

implies that

$$\mathbf{V}_{h1} = k_1^{-1/2} \mathbf{X}_{1*} (\mathbf{X}'_0 \mathbf{X}_0)^{-1} \mathbf{C}' \mathbf{F}_0^{-t} . \quad (2.29)$$

Using the result

$$\mathbf{X}'_+ \mathbf{X}_{1*} = \begin{bmatrix} \mathbf{X}'_1 & \mathbf{X}'_2 \end{bmatrix} \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{0} \end{bmatrix} = \mathbf{X}'_1 \mathbf{X}_1 = k_1 \mathbf{X}'_0 \mathbf{X}_0 , \quad (2.30)$$

the following is shown true:

$$\begin{aligned}
& \mathbf{V}'_{h_+} \mathbf{V}_{h_1} \\
&= (k_+ k_1)^{-1/2} \left[\mathbf{F}_0^{-1} \mathbf{C} (\mathbf{X}'_0 \mathbf{X}_0)^{-1} \mathbf{X}'_+ \right] \left[\mathbf{X}_{1*} (\mathbf{X}'_0 \mathbf{X}_0)^{-1} \mathbf{C}' \mathbf{F}_0^{-t} \right] \\
&= (k_1/k_+)^{1/2} \mathbf{F}_0^{-1} \mathbf{C} (\mathbf{X}'_0 \mathbf{X}_0)^{-1} (\mathbf{X}'_0 \mathbf{X}_0) (\mathbf{X}'_0 \mathbf{X}_0)^{-1} \mathbf{C}' \mathbf{F}_0^{-t} \\
&= (n_1/n_+)^{1/2} \mathbf{F}_0^{-1} \mathbf{C} (\mathbf{X}'_0 \mathbf{X}_0)^{-1} \mathbf{C}' \mathbf{F}_0^{-t} \\
&= (n_1/n_+)^{1/2} \mathbf{F}_0^{-1} \mathbf{M}_0 \mathbf{F}_0^{-t} \\
&= (n_1/n_+)^{1/2} \mathbf{F}_0^{-1} \mathbf{F}_0 \mathbf{F}'_0 \mathbf{F}_0^{-t} \\
&= (n_1/n_+)^{1/2} \mathbf{I}_a
\end{aligned} \tag{2.31}$$

Also, directly from symmetry,

$$\mathbf{V}'_{h_1} \mathbf{V}_{h_+} = \mathbf{V}'_{h_+} \mathbf{V}_{h_1} = (n_1/n_+)^{1/2} \mathbf{I}_a \quad . \tag{2.32}$$

Since $\mathbf{A}_{h_+} \mathbf{A}_{ep} = \mathbf{V}_{h_+} \mathbf{V}'_{h_+} \mathbf{V}_{ep} \mathbf{V}'_{ep} = \mathbf{0}$, the following are true:

$$\mathbf{V}'_{h_+} \mathbf{V}_{ep} = \mathbf{0} \quad , \tag{2.33}$$

and

$$\mathbf{V}'_{h_+} \mathbf{V}_{ep} = \mathbf{0} \text{ and } \mathbf{V}'_{ep} \mathbf{V}_{h_+} = \mathbf{0} \quad . \tag{2.34}$$

Similarly, since $\mathbf{A}_{ep} \mathbf{A}_{e1} = \mathbf{V}_{ep} \mathbf{V}'_{ep} \mathbf{V}_{e1} \mathbf{V}'_{e1} = \mathbf{0}$, the following are true:

$$\mathbf{V}'_{h_+} \mathbf{V}_{ep} = \mathbf{0} \tag{2.35}$$

and

$$\mathbf{V}'_{ep} \mathbf{V}_{h_+} = \mathbf{0} \quad . \tag{2.36}$$

Also,

$$\begin{aligned}
\mathbf{V}'_{ep} &= \mathbf{V}'_{ep} \mathbf{A}_{ep} \\
&= \mathbf{V}'_{ep} (\mathbf{A}_{e+} - \mathbf{A}_{e1}) \\
&= \mathbf{V}'_{ep} \mathbf{A}_{e+} - \mathbf{V}'_{ep} \mathbf{A}_{e1} \\
&= \mathbf{V}'_{ep} \mathbf{A}_{e+}
\end{aligned} \tag{2.37}$$

which implies that

$$\begin{aligned}
\mathbf{V}'_{ep}\boldsymbol{\mu}_+ &= \mathbf{V}'_{ep}\mathbf{A}_{e+}\boldsymbol{\mu}_+ \\
&= \mathbf{V}'_{ep}(\mathbf{I}_+ - \mathbf{H}_+)\mathbf{X}_+\boldsymbol{\beta} \\
&= \mathbf{V}'_{ep}(\mathbf{X}_+ - \mathbf{H}_+\mathbf{X}_+)\boldsymbol{\beta} \\
&= \mathbf{V}'_{ep}\mathbf{0}\boldsymbol{\beta} \\
&= \mathbf{0} .
\end{aligned} \tag{2.38}$$

Also,

$$\begin{aligned}
&(\mathbf{I}_{n_+} - \mathbf{A}_{ep} - \mathbf{A}_{h+})\mathbf{V}_{h1} \\
&= [\mathbf{I}_{n_+} - (\mathbf{A}_{e+} - \mathbf{A}_{e1}) - \mathbf{A}_{h+}]\mathbf{V}_{h1} \\
&= [\mathbf{I}_{n_+} - \mathbf{A}_{e+} - \mathbf{A}_{h+}]\mathbf{V}_{h1} \\
&= [\mathbf{I}_{n_+} - (\mathbf{I}_{n_+} - \mathbf{H}_+) - \mathbf{A}_{h+}]\mathbf{V}_{h1} \\
&= (\mathbf{H}_+ - \mathbf{A}_{h+})\mathbf{V}_{h1} \\
&= \mathbf{H}_+\mathbf{V}_{h1} - \mathbf{A}_{h+}\mathbf{V}_{h1} \\
&= k_+^{-1}\mathbf{X}_+(\mathbf{X}'_0\mathbf{X}_0)^{-1}\mathbf{X}'_+\mathbf{V}_{h1} - (k_1/k_+)^{1/2}\mathbf{V}_{h+} \\
&= k_+^{-1}k_1^{1/2}\mathbf{X}_+(\mathbf{X}'_0\mathbf{X}_0)^{-1}\mathbf{C}'\mathbf{F}_0^{-t} - k_+^{-1}k_1^{1/2}\mathbf{X}_+(\mathbf{X}'_0\mathbf{X}_0)^{-1}\mathbf{C}'\mathbf{F}_0^{-t} \\
&= \mathbf{0} .
\end{aligned} \tag{2.39}$$

Define the following:

$$\mathbf{y}_{h1} = \mathbf{V}'_{h1}\mathbf{y}_+ \quad a \times n_+ \times 1 \tag{2.40}$$

$$\mathbf{y}_{ep} = \mathbf{V}'_{ep}\mathbf{y}_+ \quad n_2 \times n_+ \times 1 \tag{2.41}$$

$$\mathbf{y}_{h+} = \mathbf{V}'_{h+}\mathbf{y}_+ \quad a \times n_+ \times 1 . \tag{2.42}$$

With

$$\begin{aligned}
&\begin{bmatrix} \mathbf{V}'_{h1} \\ \mathbf{V}'_{ep} \\ \mathbf{V}'_{h+} \end{bmatrix} \mathbf{I}_{n_+} [\mathbf{V}_{h1} \quad \mathbf{V}_{ep} \quad \mathbf{V}_{h+}] \\
&= \begin{bmatrix} \mathbf{I}_a & \mathbf{V}'_{h1}\mathbf{V}_{ep} & \mathbf{V}'_{h1}\mathbf{V}_{h+} \\ \mathbf{V}'_{ep}\mathbf{V}_{h1} & \mathbf{I}_{n_2} & \mathbf{V}'_{ep}\mathbf{V}_{h+} \\ \mathbf{V}'_{h+}\mathbf{V}_{h1} & \mathbf{V}'_{h+}\mathbf{V}_{ep} & \mathbf{I}_a \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{I}_a & \mathbf{V}'_{h1}\mathbf{V}_{ep} & (n_1/n_+)^{1/2}\mathbf{I}_a \\ \mathbf{V}'_{ep}\mathbf{V}_{h1} & \mathbf{I}_{n_2} & \mathbf{0} \\ (n_1/n_+)^{1/2}\mathbf{I}_a & \mathbf{0} & \mathbf{I}_a \end{bmatrix}_{(2a+n_2)} \tag{2.43}
\end{aligned}$$

and

$$\begin{bmatrix} \mathbf{V}'_{h1}\boldsymbol{\mu}_+ \\ \mathbf{V}'_{ep}\boldsymbol{\mu}_+ \\ \mathbf{V}'_{h+}\boldsymbol{\mu}_+ \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_{h1} \\ \mathbf{0} \\ \boldsymbol{\mu}_{h+} \end{bmatrix}_{(2a+n_2)\times 1}, \quad (2.44)$$

the following vector can be defined:

$$\mathbf{y}_h = \begin{bmatrix} \mathbf{V}'_{h1} \\ \mathbf{V}'_{ep} \\ \mathbf{V}'_{h+} \end{bmatrix} \mathbf{y}_+ = \begin{bmatrix} \mathbf{y}_{h1} \\ \mathbf{y}_{ep} \\ \mathbf{y}_{h+} \end{bmatrix}. \quad (2.45)$$

This is then distributed as

$$\mathbf{y}_h \sim (\mathcal{S})\mathcal{N}_{2a+n_2} \left(\begin{bmatrix} \boldsymbol{\mu}_{h1} \\ \mathbf{0} \\ \boldsymbol{\mu}_{h+} \end{bmatrix}, \sigma^2 \begin{bmatrix} \mathbf{I}_a & \mathbf{V}'_{h1}\mathbf{V}_{ep} & (n_1/n_+)^{1/2}\mathbf{I}_a \\ \mathbf{V}_{ep}\mathbf{V}_{h1} & \mathbf{I}_{n_2} & \mathbf{0} \\ (n_1/n_+)^{1/2}\mathbf{I}_a & \mathbf{0} & \mathbf{I}_a \end{bmatrix} \right). \quad (2.46)$$

Temporarily defining

$$\begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} = \sigma^2 \begin{bmatrix} \mathbf{I}_a & [\mathbf{V}'_{h1}\mathbf{V}_{ep} \quad (n_1/n_+)^{1/2}\mathbf{I}_a] \\ [\mathbf{V}_{ep}\mathbf{V}_{h1} & \mathbf{I}_{n_2} \quad \mathbf{0}] \\ (n_1/n_+)^{1/2}\mathbf{I}_a & [\mathbf{0} \quad \mathbf{I}_a] \end{bmatrix}, \quad (2.47)$$

using the properties in equation 2.39 one can write the relation:

$$\begin{aligned} \mathbf{y}_{h1} &= \mathbf{V}'_{h1}(\mathbf{A}_{ep} + \mathbf{A}_{h+})\mathbf{y}_+ + \mathbf{V}'_{h1}(\mathbf{I}_{n_+} - \mathbf{A}_{ep} - \mathbf{A}_{h+})\mathbf{y}_+ \\ &= \mathbf{V}'_{h1}(\mathbf{A}_{ep} + \mathbf{A}_{h+})\mathbf{y}_+ + \mathbf{0} \\ &= \mathbf{V}'_{h1}\mathbf{A}_{ep}\mathbf{y}_+ + \mathbf{V}'_{h1}\mathbf{A}_{h+}\mathbf{y}_+ \\ &= \mathbf{V}'_{h1}\mathbf{V}_{ep}\mathbf{y}_{ep} + \mathbf{V}'_{h1}\mathbf{V}_{h+}\mathbf{y}_{h+} \\ &= \mathbf{V}'_{h1}\mathbf{V}_{ep}\mathbf{y}_{ep} + (n_1/n_+)^{1/2}\mathbf{y}_{h+} \end{aligned} \quad (2.48)$$

with \mathbf{y}_{ep} independent of \mathbf{y}_{h+} and both distributed as described in equation 2.46. That is, the relation in equation 2.48 holds for independent \mathbf{y}_{ep} and \mathbf{y}_{h+} and

$$\mathbf{y}_{ep} \sim \mathcal{N}_{n_2}(\mathbf{0}, \sigma^2 \mathbf{I}_{n_2}) \quad (2.49)$$

$$\mathbf{y}_{h+} \sim \mathcal{N}_a(\boldsymbol{\mu}_{h+}, \sigma^2 \mathbf{I}_a). \quad (2.50)$$

Additionally,

$$\begin{aligned}
\boldsymbol{\mu}_{h_+} &= \mathbf{V}'_{h_+} \boldsymbol{\mu}_+ \\
&= k_+^{-1/2} \left[\mathbf{F}_0^{-1} \mathbf{C} (\mathbf{X}'_0 \mathbf{X}_0)^{-1} \mathbf{X}'_+ \right] \mathbf{X}_+ \boldsymbol{\beta} \\
&= k_+^{-1/2} k_+ \left[\mathbf{F}_0^{-1} \mathbf{C} (\mathbf{X}'_0 \mathbf{X}_0)^{-1} (\mathbf{X}'_0 \mathbf{X}_0) \right] \boldsymbol{\beta} \\
&= k_+^{1/2} \mathbf{F}_0^{-1} \boldsymbol{\theta} \\
&= (n_+/m)^{1/2} \mathbf{F}_0^{-1} \boldsymbol{\theta}
\end{aligned} \tag{2.51}$$

and similarly,

$$\boldsymbol{\mu}_{h_1} = (n_1/m)^{1/2} \mathbf{F}_0^{-1} \boldsymbol{\theta} . \tag{2.52}$$

These then imply the useful relation

$$\boldsymbol{\mu}_{h_1} = (n_1/n_+)^{1/2} \boldsymbol{\mu}_{h_+} . \tag{2.53}$$

For $E_s = \nu_s \widehat{\sigma}_s^2 / \sigma^2$, and $E_p = E_+ - E_1$ with $s \in \{1, +\}$, a key result needed for the results of this chapter is the distribution of E_p conditional on $\mathbf{V}'_{h_1} \mathbf{V}_{ep} \mathbf{y}_{ep}$ for the special case of $a = 1$. The following properties will help describe this distribution.

$$\begin{aligned}
\mathbf{V}'_{h_1} \mathbf{A}_{ep} \mathbf{V}_{h_1} &= \mathbf{V}'_{h_1} \mathbf{A}_{e_+} \mathbf{V}_{h_1} \\
&= \mathbf{V}'_{h_1} (\mathbf{I}_{n_+} - \mathbf{H}_+) \mathbf{V}_{h_1} \\
&= \mathbf{I}_a - \mathbf{V}'_{h_1} \mathbf{H}_+ \mathbf{V}_{h_1} \\
&= \mathbf{I}_a - k_1^{-1} \left[\mathbf{F}_0^{-1} \mathbf{C} (\mathbf{X}'_0 \mathbf{X}_0)^{-1} (\mathbf{X}_{1*})' \right] \mathbf{H}_+ \left[\mathbf{X}_{1*} (\mathbf{X}'_0 \mathbf{X}_0)^{-1} \mathbf{C}' \mathbf{F}_0^{-t} \right] \\
&= \mathbf{I}_a - (k_1/k_+) \mathbf{F}_0^{-1} \mathbf{C} (\mathbf{X}'_0 \mathbf{X}_0)^{-1} \mathbf{C}' \mathbf{F}_0^{-t} \\
&= \mathbf{I}_a - (k_1/k_+) \mathbf{F}_0^{-1} \mathbf{F}_0 \mathbf{F}'_0 \mathbf{F}_0^{-t} \\
&= \mathbf{I}_a - (k_1/k_+) \mathbf{I}_a \\
&= (n_2/n_+) \mathbf{I}_a .
\end{aligned} \tag{2.54}$$

Of course for $a = 1$, this result becomes $\mathbf{v}'_{h_1} \mathbf{A}_{ep} \mathbf{v}_{h_1} = n_2/n_+$. Using the results from equations 2.49 and 2.54, the distribution of $\mathbf{V}'_{h_1} \mathbf{V}_{ep} \mathbf{y}_{ep}$ can then be stated:

$$\mathbf{V}'_{h_1} \mathbf{V}_{ep} \mathbf{y}_{ep} \sim \mathcal{N}_a \left[\mathbf{0}, (n_2/n_+) \sigma^2 \mathbf{I}_a \right] . \tag{2.55}$$

Or, for $a = 1$,

$$\mathbf{v}'_{h_1} \mathbf{V}_{ep} \mathbf{y}_{ep} \sim \mathcal{N} \left[0, (n_2/n_+) \sigma^2 \right] . \tag{2.56}$$

A useful result is the distribution of \mathbf{y}_{ep} conditional on $\mathbf{V}'_{h1}\mathbf{V}_{ep}\mathbf{y}_{ep}$ for the $a = 1$ case.

For $a = 1$, their joint distribution is

$$\begin{bmatrix} \mathbf{y}_{ep} \\ \mathbf{v}'_{h1}\mathbf{V}_{ep}\mathbf{y}_{ep} \end{bmatrix} = \begin{bmatrix} \mathbf{I}_{n_2} \\ \mathbf{v}'_{h1}\mathbf{V}_{ep} \end{bmatrix} \mathbf{y}_{ep} \sim (S)\mathcal{N}_{n_2+1} \left(\begin{bmatrix} \mathbf{0} \\ 0 \end{bmatrix}, \sigma^2 \begin{bmatrix} \mathbf{I}_{n_2} & \mathbf{V}'_{ep}\mathbf{v}_{h1} \\ \mathbf{v}'_{h1}\mathbf{V}_{ep} & (n_2/n_+) \end{bmatrix} \right). \quad (2.57)$$

Conditional Gaussian theory (see Muller and Stewart, 2006; Chapter 8) then implies for $a = 1$, $\boldsymbol{\mu}_{e.h} = (n_+/n_2)t_0\mathbf{V}'_{ep}\mathbf{v}_{h1}$, and $\boldsymbol{\Sigma}_{e.h} = \sigma^2[\mathbf{I}_{n_2} - (n_+/n_2)\mathbf{V}'_{ep}\mathbf{A}_{h1}\mathbf{V}_{ep}]$, the distribution of \mathbf{y}_{ep} conditional on $\mathbf{v}'_{h1}\mathbf{V}_{ep}\mathbf{y}_{ep}$ is

$$\mathbf{y}_{e.h} = \mathbf{y}_{ep} | (\mathbf{v}'_{h1}\mathbf{V}_{ep}\mathbf{y}_{ep} = t_0) \sim (S)\mathcal{N}_{n_2}(\boldsymbol{\mu}_{e.h}, \boldsymbol{\Sigma}_{e.h}) . \quad (2.58)$$

For $E_p = \mathbf{y}'_{ep}\mathbf{y}_{ep}/\sigma^2$, I can now state in a corollary the distribution of E_p conditional on $\mathbf{V}'_{h1}\mathbf{V}_{ep}\mathbf{y}_{ep}$ for the special case of $a = 1$.

Corollary 2.1 For $a = 1$ and $E_p = \mathbf{y}'_{ep}\mathbf{y}_{ep}/\sigma^2$, the distribution of E_p conditional on $\mathbf{v}'_{h1}\mathbf{V}_{ep}\mathbf{y}_{ep} = t_0$ is

$$E_p | (\mathbf{v}'_{h1}\mathbf{V}_{ep}\mathbf{y}_{ep} = t_0) = X_p + (n_+/n_2)(t_0/\sigma)^2 \quad (2.59)$$

where X_p is a central χ^2 distributed variable with $n_2 - 1$ degrees of freedom.

A proof of Corollary 2.1 is in Appendix A.

2.3 THE IPIA PROCEDURE AND PROPERTIES

Table 2.4: *General procedure*

Step 1a : Specify $\alpha_t, P_t, \mathbf{X}_0$, hypotheses, θ_1 , and σ_0^2
1b : Solve for first stage sample size (n_1)
Step 2 : Collect first n_1 observations
Step 3 : Solve for $N_+ = n_+$, critical values $f_l(n_+)$, $f_u(n_+)$, and $f_+(n_+)$, and F_1
Step 4 : Decide:
If $F_1 < f_l$ then STOP, ACCEPT H_0
If $F_1 \geq f_u$ then STOP, REJECT H_0
If $f_l \leq F_1 < f_u$ then take $n_2 = n_+ - n_1$ additional observations
Step 5 : Solve for F_+
Step 6 : Decide:
If $F_+ < f_+$ then ACCEPT H_0
If $F_+ \geq f_+$ then REJECT H_0

Table 2.4 outlines the general procedure for the IPIA model. The order of the steps matters in specifying the distributions. The above sequence seems the most sensible.

The value of the internal pilot sample size, n_1 , should be made at the design stage of the study. The choice is important since lower values give more uncertain estimates of σ^2 while higher values reduce possible savings in sample size. For traditional internal pilot designs, Birkett and Day (1994) recommended at least 20 degrees of freedom in the first stage sample; however, Coffey and Muller (1999) showed that this scenario can still produce type I error rate inflation depending on the design and noted the importance of examining properties of specific study designs. A default in literature seems to be taking a designated fraction of the sample size from fixed sample equations using the best guess for the variance at the planning stage, i.e., $\pi \cdot n_0$ for $0 < \pi \leq 1$ and n_0 determined from σ_0^2 . A default choice for π seems to be $1/2$; that is, the size of the first sample is half of the fixed sample study sample size based on σ_0^2 . For this chapter, a choice of π as close to $1/2$ as possible will be chosen. The effect of the initial sample size choice in the IPIA design will be considered as a factor in examples in Chapter 3.

Calculation of the three critical values for the study, $f_l(n_+)$, $f_u(n_+)$, and $f_+(n_+)$, must be done following rules pre-specified in the study protocol. The critical values may depend on n_+ , the realized value of N_+ ; however, when it is clear, they will be referred to as f_l , f_u , and f_+ . Ideally, they should be chosen in a way that controls the type I error rate while having good power and expected sample size properties. The theory developed here optionally allows for stopping under the null at the interim analysis if $F_1 < f_l$, where f_l is the first stage lower critical value. This can cause a great reduction in expected sample size when the effect size is near the null value by allowing the study to stop for a "lost cause". If no early stopping ability under the null is desired, then $f_l = 0$ for all $n_+ \neq n_1$. In all cases $f_l = f_u$ when $n_+ = n_1$, which guarantees stopping for acceptance or rejection of the null. More detailed exploration and comparison of selection methods will be undertaken in Chapter 3 of this dissertation.

The sample size re-estimation rule will determine the distribution of N_+ . It is an important consideration in the design affecting type I error rate, power, and expected sample size. Like internal pilot designs, the sample size for IPIA designs is determined by using the updated variance estimate at the interim stage to recalculate the estimated sample size need to achieve target power in the final test. The procedure takes advantage of the monotone relationship between continuous $\hat{\sigma}_1^2$ and discrete N_+ . In order to determine the distribution of sample size, one solves the cumulative distribution function for possible values of N_+ , that is $\Pr\{N_+ \leq n\}$, by determining cut-off points based on the first stage variance estimate. For a particular value n_+ of random N_+ , one solves for scaled noncentrality $\lambda(n_+)$ that satisfies

$$P_t = 1 - F_{\chi^2}[f_{\text{crit}}; a, \lambda(n_+)] \quad (2.60)$$

or

$$P_t = 1 - F_F[f_{\text{crit}}; a, \nu_+, \lambda(n_+)] \quad (2.61)$$

where $f_{\text{crit}} = F_{\chi^2}^{-1}(1 - \alpha_t; a)$ or $F_F^{-1}(1 - \alpha_t; a, \nu_+)$ depending on whether or not large

sample distributional assumptions are used. The scaled noncentrality $\lambda(n_+)$ here represents the minimum value that would lead to a final sample size of n_+ . Since the effect of interest, θ_1 , is used at the planning stage, for $\delta(n_+) = \theta_1' \mathbf{M}_+^{-1} \theta_1$ the following hold:

$$\lambda(n_+) = \delta(n_+)/\sigma^2(n_+) \quad (2.62)$$

or

$$\sigma^2(n_+) = \delta(n_+)/\lambda(n_+). \quad (2.63)$$

Here, $\sigma^2(n_+)$ designates the largest value of $\hat{\sigma}_1^2$ that would produce $N_+ = n_+$. Therefore, since $\nu_1 \hat{\sigma}_1^2 \sim \sigma^2 W$ for $W \sim \chi^2(\nu_1)$,

$$\begin{aligned} \Pr\{N_+ \leq n_+\} &= \Pr\{\hat{\sigma}_1^2 \leq \sigma^2(n_+)\} \\ &= \Pr\{W \leq \nu_1 \sigma^2(n_+)/\sigma^2\} \\ &= \Pr\{W \leq \nu_1 \delta(n_+)/[\sigma^2 \lambda(n_+)]\}. \end{aligned} \quad (2.64)$$

The discreteness of sample size implies

$$\Pr\{N_+ = n_+\} = \Pr\{N_+ \leq n_+\} - \Pr\{N_+ \leq n_+ - m\}. \quad (2.65)$$

When restrictions are given for minimum or maximum sample size, the tail probabilities are collapsed into the smallest or largest allowable values, respectively.

A key result of this process is the determination of cut-off points that determine a range into which continuous $\hat{\sigma}_1^2$ must have fallen in order for a given final sample size to occur.

Define $q_1(n_+)$ and $q_2(n_+)$ to be the values such that

$$N_+ = n_+ \Leftrightarrow q_1(n_+) < \nu_1 \hat{\sigma}_1^2 / \sigma^2 \leq q_2(n_+), \quad (2.66)$$

which in turn implies that

$$\Pr\{N_+ = n_+\} = F_{\chi^2}[q_2(n_+); \nu_1] - F_{\chi^2}[q_1(n_+); \nu_1]. \quad (2.67)$$

The cut off points determine the probabilities for discrete values of N_+ and hence describe the variable's distribution. When it is unambiguous, q_1 and q_2 are used for $q_1(n_+)$ and $q_2(n_+)$.

One small difference between this technique and that used for sample size calculations in group sequential methods is the lack of the sample size inflation factor used to account for the drop in power from multiple looks. While its use could be integrated into the method without much difficulty, it would have very little effect on this two-stage design. For example, using O'Brien-Fleming bounds for $P_t = 0.9$ and $\alpha_t = 0.05$ calls for a sample size inflation factor of 1.007 (Jennison and Turnbull, 2000; Chapter 2). Additionally, changing the P_t and α_t does not change this factor much. Due to its lack of significance, I decided to simplify the method by leaving it out for now; it can always be added if deemed important.

2.4 KEY ANALYTIC RESULTS FOR PROCEDURE

The results in this paper are developed to test $H_0 : \theta = 0$ with $\theta = \mathbf{C}\boldsymbol{\beta}$ for the case where contrast matrix \mathbf{C} has only one row, namely, $a = 1$. Model designs falling under this restriction include one and two group mean comparisons, tests for interaction, and other designs of interest. Key results in this section will be specific to this design restriction.

In order to compute power (and hence, type I error rate) for the study design, a joint distribution of the two stage test statistics if necessary. Since critical values and denominator degrees of freedom depend on sample sizes, I derive the joint CDF conditional on an N_+ value, or $F_{F_1, F_+ | N_+}(f_1, f_+)$. By using the law of total probability and summing over possible values, this result leads to calculation of unconditional power.

Conditional on N_+ , $E_1 = \nu_1 \hat{\sigma}_1^2 / \sigma^2 \sim \chi_T^2[\nu_1; q_1(n_+), q_2(n_+)]$ (Coffey & Muller, 1999).

The following theorem gives a convenient expression for the conditional joint CDF of the test statistics for the $a = 1$ case.

Theorem 2.1 Related results and definitions can be found in section 2.2.3. Define $b(t_e, t_h) = t_h^2 / (\sigma^2 c_+) - t_e$, $d(t_p) = n_+ t_p^2 / (n_2 \sigma^2)$, $h(t_e, t_h) = \sigma(c_1 t_e)^{1/2} - (n_1 / n_+)^{1/2} t_h$,

$l(t_e, t_h) = -\sigma(c_1 t_e)^{1/2} - (n_1/n_+)^{1/2} t_h$, and $p_{n_+} = F_{\chi^2}(q_2; \nu_1) - F_{\chi^2}(q_1; \nu_1)$. For $a = 1$,

$$\begin{aligned} & F_{F_1, F_+ | N_+}(f_1, f_+) \\ &= p_{n_+}^{-1} \int_{q_1}^{q_2} f_{\chi^2}(t_e; \nu_1) \int_{-\infty}^{\infty} f_{\mathcal{N}}(t_h; \mu_{h+}, \sigma^2) \int_{l(t_e, t_h)}^{h(t_e, t_h)} f_{\mathcal{N}}[t_p; \mathbf{0}, (n_2/n_+)\sigma^2] \\ & \quad \times \{1 - F_{\chi^2}[b(t_e, t_h) - d(t_p); n_2 - 1]\} dt_p dt_h dt_e \end{aligned} \quad (2.68)$$

A proof of Theorem 2.1 is in Appendix A.

A needed result when early futility stopping is allowed follows directly from Theorem 2.1 as follows

$$\Pr\{f_l \leq F_1 \leq f_u, F_+ \leq f_+ | N_+ = n_+\} = F_{F_1, F_+ | N_+}(f_u, f_+) - F_{F_1, F_+ | N_+}(f_l, f_+). \quad (2.69)$$

A special case distribution is the condition joint distribution of the test statistics when $n_2 = 1$. This result would be needed for single group hypothesis tests and is a simplification of Theorem 2.1.

Theorem 2.2. Related results and definitions can be found in section 2.2.3. For $a = 1$ and $n_2 = 1$, define $b(t_e) = n_+ c_1 t_e$ and $p_{n_+}^{-1} = F_{\chi^2}(q_2; \nu_1) - F_{\chi^2}(q_1; \nu_1)$, then

$$\begin{aligned} & F_{F_1, F_+ | N_+}(f_1, f_+) \\ &= p_{n_+}^{-1} \int_{q_1}^{q_2} f_{\chi^2}(t_e; \nu_1) \int_0^{b(t_e)} f_{\chi^2}(t_p; n_2) \\ & \quad \times F_{\chi^2}\{\min[(b(t_e) - t_p)/n_1, c_+(t_e + t_p)], 1, \lambda_+\} dt_h dt_e \end{aligned} \quad (2.70)$$

A proof of Theorem 2.2 is in Appendix A.

Another distribution important to power and expected size calculations is the CDF of the first test statistic, conditional on N_+ , i.e., $F_{F_1 | N_+}(f_1)$. This result is not new and can be found in Coffey and Muller (1999) who considered it for use in internal pilots for the case of $N_+ = n_1$. In the context of the IPIA model it has greater importance due to early stopping possibilities and is presented here to use in the forthcoming equations for power and expected sample size.

Theorem 2.3 The conditional CDF of the first test statistic can be written

$$F_{F_1|N_+}(f_1) = \int_{q_1}^{q_2} \frac{F_{\chi^2}(c_1 t_1; a, \lambda_1) f_{\chi^2}(t_1; \nu_1)}{F_{\chi^2}(q_2; \nu_1) - F_{\chi^2}(q_1; \nu_1)} dt_1 . \quad (2.71)$$

A proof of Theorem 2.3 is in Appendix A.

The following corollary adapts Theorem 2.3 to solve for the probability of the test continuing to the second stage conditional on $N_+ = n_+$ when futility stopping is possible.

Corollary 2.2

$$\begin{aligned} & \Pr\left\{f_l \leq F_1 < f_u \mid N_+ = n_+\right\} \\ &= \int_{q_1}^{q_2} \frac{[F_{\chi^2}(c_u t_1; a, \lambda_1) - F_{\chi^2}(c_l t_1; a, \lambda_1)] f_{\chi^2}(t_1; \nu_1)}{F_{\chi^2}(q_2; \nu_1) - F_{\chi^2}(q_1; \nu_1)} dt_1 \end{aligned} \quad (2.72)$$

The proof is similar to proof of Theorem 2.3, in Appendix A.

The above results can then be used to calculate exact expressions for the power, type I error rate (power under the null hypothesis), and expected sample size in the study. The values change with design parameters and are valuable knowledge in study planning. *The following theorem gives the formula for unconditional power.*

Theorem 2.4 An expression for unconditional power, P_w , can be written

$$\begin{aligned} P_w = 1 - \sum_{\{N_+=n_+\}} [F_{\chi^2}(q_2; \nu_1) - F_{\chi^2}(q_1; \nu_1)] & \left\{ F_{F_1|N_+}(f_l(n_+)) + \right. \\ & \left. \Pr\left[f_l(n_+) \leq F_1 < f_u(n_+), F_+ < f_+(n_+) \mid N_+ = n_+\right] \right\} . \end{aligned} \quad (2.73)$$

A proof of Theorem 2.4 is in Appendix A.

The results in this section can also be applied to calculate an expected sample size formula for a study design in the following form.

Theorem 2.5 Let $N_w =$ total sample size taken in study, that is,

$$N_w = \begin{cases} n_1 & \text{if study stopped after first stage} \\ N_+ & \text{otherwise} \end{cases}$$

then

$$E(N_w) = n_1 + \sum_{\{N_+=n_+\}} n_2 [F_{\chi^2}(q_2; \nu_1) - F_{\chi^2}(q_1; \nu_1)] \Pr[f_l \leq F_1 < f_u | N_+ = n_+] \quad (2.74)$$

A proof of Theorem 2.5 is in Appendix A.

2.5 EXAMPLES

2.5.1 Motivation for Examples

Two example study designs are considered in this chapter and revisited in Chapter 3, where design strategies are evaluated. The designs are both two group comparisons, but have different sample size needs. There are three main purposes of the examples in this chapter. First, to compare the numeric results using exact theory to simulations in order to justify the computational algorithms utilized in calculations and as an additional check on the accuracy of the theory. Second, to examine the properties of the proposed internal pilot with interim analysis (IPIA) procedure and how they compare to properties of special cases such as fixed sample, internal pilot (IP), or two-stage group sequential (GS) procedures. To facilitate this purpose, I use a naive, but common approach to critical value selection and study design. Properties to be examined include type I error rate, power, and expected sample sizes under various scenarios. Finally, a comparison of the two examples will allow for an illustration of properties and designs that are sensitive to study sample size.

The fixed sample, IP and two-stage GS designs considered for the examples in this chapter are all special cases within the general IPIA framework introduced. The IP design does not allow for early stopping at the interim power analysis (special case of IPIA: $f_l = 0$, $f_u = \infty$) and in the design used here, it is assumed that sample size can be reduced from the pre-planned level ($n_{+,min} = n_1$). The two-stage GS design, on the other hand, allows for early stopping at the interim analysis, but does not allow for a change to the preplanned maximum sample size, or: $\Pr\{N_+ = n_0\} = 1$. The fixed sample approach can be seen as a

special case combining the restrictions of the IP and GS designs ($f_l = 0$, $f_u = \infty$, and $\Pr\{N_+ = n_0\} = 1$). The IPIA design combines the features of the IP and two-stage GS designs by allowing for stopping at the interim stage as well as allowing for a change in maximum sample size used when a study is to be continued.

Table 2.5: *Two-stage designs*

		Early Stopping	
		Yes	No
SSR	Yes	IPIA	Int Pilot
	No	Grp Seq	Fixed Sample

In addition to stopping at the interim stage for efficacy, both the GS and the IPIA procedures can have possible early stopping at the interim stage for futility. Hence, they will be considered here under both scenarios. No futility stopping implies that the lower first stage critical value, f_l , is 0. For futility stopping in this chapter, I will use a simple p-value cut point of 0.85. In other words, if the p-value for the first stage hypothesis test is greater than 0.85 then the study will stop and conclude that the alternative hypothesis is not supported. In reality, this is not an ideal approach since the first stage may contain only a small fraction of the needed information of the study (especially for high true variance values) and so may not be very informative in some cases. It is used here for simplicity in order to portray the characteristics of the procedures.

Critical values used in the examples will be calculated as follows. Critical values for the fixed sample, group sequential, and IPIA designs will be based on the standard Gaussian distribution. This is to show the consequences of not accounting for the use of variance estimates in the test statistics for the small and moderate studies examined. The fixed sample result using the t distribution will also be included since it will exactly achieve the target type I error rate. The critical values for the internal pilot design will only be solved with the unadjusted t distribution since it is the method described by literature. O'Brien-Fleming stopping rule bounds determined by the design and information fraction at the interim

analysis will be used for the group sequential and IPIA (early stopping) designs. These bounds are designed to allow for conservative early stopping while adjusting the final critical value for type I error rate inflation due to multiple testing.

Due to misalignment of test statistic and critical value distributions and a biased variance value used in sample size re-estimation designs, the unadjusted selection for a final critical values will likely cause type I error rate inflation for most designs considered. They are used here in order to best compare the procedures and also to notice the magnitudes of such inflations. Strategies that better control the type I error rate and maintain power while minimizing expected sample size will be considered in Chapter 3 when these examples are revisited. The importance here is to illustrate design characteristics and comparisons.

In total seven design procedures will be considered: fixed sample (Z and t), IP, two-stage GS with and without futility stopping, and IPIA with and without futility stopping. Type I error rate, power, and expected sample size will be calculated for each of these procedures for true variance values determined by $\gamma \in \{0.5, 0.75, 1, 1.5, 2\}$ where $\gamma = \sigma^2/\sigma_0^2$. I will use the sampling fraction $\pi = 0.5$ and $n_{+, \max} = \infty$. Since the expected sample size of procedures with early stopping (GS and IPIA with and without futility stopping) depends on the true effect size θ , I will calculate values under three conditions: under the null ($\theta = 0$), under the alternative of interest ($\theta = \theta_1$), and under a true effect twice that of the effect size of interest ($\theta = 2\theta_1$).

2.5.2 Computational Methods

All programs for the examples were written in SAS/IML (SAS Institute, 2004). Most of the computation for the examples utilizes the exact theory developed in this chapter. The exceptions are the fixed sample and internal pilot designs. The fixed results could be directly calculated using standard distribution functions. The internal pilot calculations were easily obtained utilizing exact internal pilot theory from the freely available GLUMIP 2.0 (Kairalla

et al., 2007) software package. All other results came from use of the exact theory, including the two-stage group sequential designs, which are a special case.

Stopping bound computation utilized the SEQSCALE function and the numeric integrations utilized the QUAD function within SAS/IML. To avoid numerical instability of the calculated integrals, computation was performed using quantile transformations (Glueck and Muller, 2001) of the distributions derived in Section 2.4. For illustration, let $p = F_{\chi^2}(t; \nu)$. Then the integral transformation using $t = F_{\chi^2}^{-1}(p; \nu)$ and $dp = f_{\chi^2}(t; \nu)dt$ allows for integration using finite bounds with better behaved integrands.

Simulations were conducted for a limited set of cases in order to check the accuracy of the programming and numerical algorithms, provide an additional check on the analytical derivations, and to compare the speed of calculation using the two methods. Using a subset of a dozen cases from both examples over a range of conditions, I conducted simulation with 1,000,000 replications per case. All programs were run using an Intel Xeon 3.2 GHz processor. For each of the cases considered, the analytically calculated values were within two standard deviations of the simulated values.

The comparison programs were each run in groups of three cases corresponding to variance values of $\gamma \in \{0.5, 1, 2\}$ with $\gamma = \sigma^2/\sigma_0^2$. Runs were made under the null hypothesis ($\theta = 0$) and assuming the effect of interest ($\theta = \theta_1$) for the IPIA design without futility for both the moderate (Example 2.1, Section 2.5.3) and small sample examples (Example 2.2, Section 2.5.4) considered in this chapter. Timing results are detailed in Table 2.6 below.

Table 2.6: *Simulation and calculation times (minutes)*

	$\theta = 0$ (null)		$\theta = \theta_1$ (alt)	
	Sim	Calc	Sim	Calc
Ex. 2.1	176.3	4.6	171.9	7.1
Ex. 2.2	124.2	1.8	124.5	2.8

Clearly from the results above, the analytic calculations using the exact theory are much faster than the simulations for all cases. While little effort has been made thus far to improve the computational speed of the programs, the simulations took between 24-70 times more computation time depending on the case. It is of note that the alternative case results took longer for the analytical calculations but made little difference in the simulations. This is most likely due to the various noncentralities that come into play in such circumstances. Also, for both methods, the timing is larger for the larger design (Example 2.1). The proportional increase is greater for the analytical results due to the increase in conditional cases that must be considered and calculated. The timing benefits remain clear however with the worse case still being over 20 times more efficient.

2.5.3 Example 2.1 Results

The design parameters for Example 2.1 are summarized in Table 2.7.

Table 2.7: *Design parameters for Example 2.1*

α_t	P_t	θ_1	σ_0^2	n_0	n_1
0.05	0.9	1	2	86	44

Example 2.1, a study design of moderate size, was considered by Wittes and Brittain (1990) and Coffey and Muller (1999) in an internal pilot framework. Values are calculated analytically for type I error rate, power, and expected sample size under the design conditions described in Table 2.7 for Example 2.1.

Table 2.8: *Type I error rates $\times 100$ for Example 2.1*

γ	Fixed		IP	Group Sequential		IPIA	
	Z	t		w/o Futility	w/ Futility	w/o Futility	w/ Futility
0.50	5.3	5.0	5.2	5.5	5.4	5.8	5.8
0.75	5.3	5.0	5.4	5.5	5.4	6.0	6.0
1.00	5.3	5.0	5.3	5.5	5.4	5.9	5.9
1.50	5.3	5.0	5.2	5.5	5.4	5.6	5.4
2.00	5.3	5.0	5.2	5.5	5.4	5.4	5.1

Table 2.8 displays the values for type I error rate for each of the seven designs described in Section 2.5.1: Fixed sample (Z and t), IP, two-stage GS with and without futility stopping,

and IPIA with and without futility stopping. In order to see results as a function of variance, values are calculated for true variances corresponding to values of $\gamma = \sigma^2/\sigma_0^2$ with $\gamma \in \{0.5, 0.75, 1, 1.5, 2\}$.

For the fixed sample design, type I error rate is somewhat inflated by a constant amount across γ when the standard Gaussian distribution is used for critical value determination and is controlled at the target level when the t distribution is used. In the IP design, mild type I error rate inflation occurs due to downward bias in variance estimate used. The magnitude of inflation is shown to depend on true variance value. Due to the use of Gaussian critical values, the GS designs also have type I error rate inflation. The inflation for the GS designs is constant across γ since no sample size re-estimation occurs and noncentrality is zero under the null. The IPIA designs, which combine early stopping ability with sample size re-estimation, have moderate type I error rate inflation caused by both variance estimate bias and the use of Gaussian critical values. For both GS and IPIA, allowing for early stopping for futility causes a small reduction in the type I error rate. For this example, the magnitude of type I error rate is comparable between the Fixed, IP, and GS methods and the IPIA method has an somewhat increased level of inflation.

Table 2.9: *Power $\times 100$ for Example 2.1*

γ	Fixed		IP	Group Sequential		IPIA	
	Z	t		w/o Futility	w/ Futility	w/o Futility	w/ Futility
0.50	99.6	99.6	92.9	99.6	99.5	93.0	93.0
0.75	96.5	96.3	90.6	96.4	96.2	90.3	90.2
1.00	90.5	90.0	90.0	90.3	90.0	89.8	89.5
1.50	76.3	75.4	89.4	76.1	75.7	89.4	88.0
2.00	64.1	63.0	89.2	63.9	63.4	89.1	86.5

Table 2.9 displays the values for unconditional power for the seven designs. Power for both fixed sample designs is sensitive to the true variance value. With a target level of 90%, a fixed sample study can be significantly over or under powered depending on the true variance regardless of the critical value determination method employed. Power for the considered GS designs is also highly dependent on the true variance value, with power levels

very similar to those for the fixed sample design. In the IP design, power is greatly stabilized due to the variance estimate based sample size re-estimation at first stage. The IPIA designs also have very stable power similar to the IP design due to the sample size re-estimation analysis at the first stage. The GS and IPIA designs with first stage futility stopping have power at slightly lower levels than their counterparts without futility stopping. For power in this example, the IP and IPIA designs greatly achieve the target rate while the two-stage GS and fixed sample designs are shown to be vulnerable to misspecification of the variance, a nuisance parameter, at the planning stage.

Table 2.10: $E(N_w)$ for Example 2.1: fixed, IP, and GS

γ	Fixed	IP	GS (no futility)			GS (futility)		
			$\theta = 0$	$\theta = \theta_1$	$\theta = 2\theta_1$	$\theta = 0$	$\theta = \theta_1$	$\theta = 2\theta_1$
0.50	86	48.1	85.6	56.3	44.0	79.4	56.3	44.0
0.75	86	66.2	85.6	65.7	44.2	79.4	65.7	44.2
1.00	86	87.0	85.6	71.3	45.3	79.4	70.9	45.3
1.50	86	129.0	85.6	77.1	50.3	79.4	76.1	50.3
2.00	86	171.0	85.6	79.8	56.3	79.4	78.2	56.3

Table 2.10 displays the values for expected sample size for the fixed sample, IP, and GS with and without futility stopping designs. For the GS designs, expected sample sizes are calculated assuming the null hypothesis ($\theta = 0$), the alternative of interest ($\theta = \theta_1$), and assuming a true effect size twice the effect of interest ($\theta = 2\theta_1$).

Under controlled conditions, the sample size for the fixed sample design is always the preplanned sample size, 86. As would be expected, the expected sample size for the IP design is dependent on the true variance due to sample estimate based sample size re-estimation at the first stage. It achieves an expected savings in sample size for variance values lower than the value assumed at the planning stage and rises above that of the fixed sample design as it accounts for larger true variance values by increasing the estimated sample size need at the internal pilot stage. Under the null hypothesis, the expected sample size for the GS designs are constant over γ values. This is because no variance value based sample size re-estimation takes place at first stage and true noncentrality is zero. The small

departure from the fixed design sample size in the GS design without futility stopping is due to the small chance of falsely stopping for efficacy at the first stage. The GS designs allowing futility stopping at the first stage causes an across the board drop in expected sample size under the null due the probability of correctly stopping early for futility. Under the alternative of interest ($\theta = \theta_1$), the expected sample sizes for the GS designs are noticeable lower than the fixed design sample sizes due to possible early stopping for efficacy at first stage. This demonstrates the clear sample size benefits of the GS designs compared to single analysis, fixed sample designs. The effect diminishes as variance increases due to the lowered power of the first test with decreasing noncentrality of the test statistic. When futility stopping is allowed, the expected sample size for the GS design decreases slightly as the probability of false futility stopping at the first stage analysis is introduced. For an effect of twice the alternative of interest ($\theta = 2\theta_1$), the GS designs offer significant expected sample size reduction at all γ considered. The effect diminishes somewhat as first stage power decreases for increasing variance. There is very little difference in the two GS designs considered under this condition as the chance of first stage futility stopping is very small for $\theta = 2\theta_1$.

Table 2.11: $E(N_w)$ for Example 2.1: IPIA

γ	IPIA (no futility)			IPIA (futility)		
	$\theta = 0$	$\theta = \theta_1$	$\theta = 2\theta_1$	$\theta = 0$	$\theta = \theta_1$	$\theta = 2\theta_1$
0.50	47.2	44.8	44.0	46.6	44.8	44.0
0.75	64.0	54.3	44.2	60.8	54.3	44.2
1.00	84.7	73.4	47.1	78.3	73.0	47.1
1.50	126.8	119.8	78.3	114.0	117.7	78.3
2.00	169.0	165.5	133.8	149.7	160.6	134.0

Table 2.11 displays the values for expected sample size for the IPIA designs with and without futility stopping. Similar to the IP design, IPIA expected samples sizes are lower than the fixed design sample size for low true variance and increase with the true variance as the first stage sample size re-estimation requires larger second stage samples on average. Under the null hypothesis ($\theta = 0$), the IPIA design without futility stopping has very similar

sample size values to the IP design since stopping at the first stage for efficacy is rare here. The null case IPIA design with possible futility stopping causes a drop in expected sample size for all variance values when compared the design without futility stopping due to the chance of a correct decision to accept the null and stop at the first stage analysis. Under the alternative of interest, ($\theta = \theta_1$), the expected sample sizes for the IPIA designs are noticeably lower than the fixed sample design for variance values at the preplanned value or lower. The expected sample sizes rise with γ due to the increased sample size needs detected at the first stage analyses to protect study power. In this case, early stopping (GS-like) sample size benefits are offset by the sample size recalculation procedure (power protecting IP characteristic). Expected sample size under the alternative of interest is slightly lower in the IPIA design allowing futility stopping due to the possibility of false futility stops at first stage. This chance increases with γ due to the naive p-value based futility stopping rules used to calculate the first stage futility critical value. For an effect of twice the alternative of interest ($\theta = 2\theta_1$), the IPIA designs offer significant expected sample size reduction due the large chance of early efficacy stopping in the first stage. The effect diminishes as increasing variance calls for more sample size in the second stage and decreases first stage power. There is very little difference in the two IPIA designs considered under $\theta = 2\theta_1$ as the chance of futility stopping is very small for the large effect size.

2.5.4 Example 2.2 Results

The design parameters for Example 2.2 are summarized in Table 2.12.

Table 2.12: *Design parameters for Example 2.2*

α_t	P_t	θ_1	σ_0^2	n_0	n_1
0.05	0.9	1.6	1	20	10

Example 2.2 is a smaller study than Example 2.1 and was considered by Coffey and Muller (1999) in order to explore small sample properties for internal pilots. It will be useful to

describe characteristics of the study designs that are sensitive to the planning stage sample size estimate for the study. Values are calculated analytically for type I error rate, power, and expected sample size under the design conditions described in Table 2.13 for Example 2.2.

Table 2.13: *Type I error rates $\times 100$ for Example 2.2*

γ	Fixed		IP	Group Sequential		IPIA	
	Z	t		w/o Futility	w/ Futility	w/o Futility	w/ Futility
0.50	6.6	5.0	5.5	7.8	7.7	8.8	8.8
0.75	6.6	5.0	6.2	7.8	7.7	9.3	9.3
1.00	6.6	5.0	6.5	7.8	7.7	9.6	9.5
1.50	6.6	5.0	6.5	7.8	7.7	9.4	9.2
2.00	6.6	5.0	6.2	7.8	7.7	8.8	8.5

Table 2.13 displays the values for type I error rate for the seven designs described in Section 2.5.1. Values are calculated for true variances corresponding to values of $\gamma = \sigma^2/\sigma_0^2$ with $\gamma \in \{0.5, 0.75, 1, 1.5, 2\}$.

The type I error rate for the standard Gaussian fixed sample design here is significantly inflated. This is due to using standard Gaussian critical values that do not take into account the uncertainty of the variance estimate. Like the previous example, using the correct t based distribution exactly achieves the target rate. Like the Gaussian fixed design, the GS designs also have type I error rate inflation due to large sample nature of the critical value selection methods. The inflation for the fixed and GS methods is constant over true variance values. In the IP and IPIA designs, type I error rate inflation again occurs as a function of variance. The IPIA method has the greatest amount of inflation due to both the biased variance estimate being used as well as the Gaussian critical values. For both GS and IPIA designs, allowing for early stopping for futility causes a small reduction in the type I error rate.

For Example 2.2, type I error rates have similar trends to values calculated in Example 2.1, but with differing magnitudes. All magnitudes of type I error rate inflation are magnified here by the small nature of the study compared to the moderately larger study

examined in Example 2.1. Additionally, the magnitude difference for the IPIA designs is the greatest due to being vulnerable to both sources of inflation.

Table 2.14: *Power $\times 100$ for Example 2.2*

γ	Fixed		IP	Group Sequential		IPIA	
	Z	t		w/o Futility	w/ Futility	w/o Futility	w/ Futility
0.50	99.8	99.8	96.1	99.8	99.8	96.7	96.7
0.75	97.4	97.4	93.2	98.2	98.1	93.6	93.6
1.00	94.1	92.2	91.3	94.2	94.0	91.7	91.5
1.50	82.6	78.9	88.8	82.9	82.6	90.5	88.6
2.00	71.5	66.8	87.3	72.0	71.6	88.2	86.3

Table 2.14 displays the values for unconditional power for Example 2.2. As in Example 2.1, power for fixed sample and GS designs is very sensitive to the true variance value. With a target level of 90%, these designs can be significantly over or under powered if the variance is misspecified at planning. In the IP and IPIA designs, power is again greatly stabilized due to the sample size re-estimation at the first stage analysis based on the sample variance estimate. For power in this example, the IP and IPIA designs greatly achieve the target rate while the two-stage GS and fixed sample designs are shown to be vulnerable to misspecification of the variance, a nuisance parameter, at the study planning stage.

A comparison of power in Examples 2.1 and 2.2 shows that the size of a study does not have a clear effect on the power functions for these study designs in general or as a function of $\gamma = \sigma^2/\sigma_0^2$. In other words, the planned size of the study can not appreciably relieve the dangers of over or under powering a study based on nuisance parameter misspecification at the planning stage for the examined fixed sample and GS designs. The results show that sample size re-estimation procedures are effective ways to help stabilize study power so that

important questions of interest can be explored accurately and dependably.

Table 2.15: $E(N_w)$ for Example 2.2: fixed, IP, and GS

γ	Fixed	IP	GS (no futility)			GS (futility)		
			$\theta = 0$	$\theta = \theta_1$	$\theta = 2\theta_1$	$\theta = 0$	$\theta = \theta_1$	$\theta = 2\theta_1$
0.50	20	12.3	19.8	12.4	10.0	18.3	12.4	10.0
0.75	20	15.9	19.8	14.3	10.1	18.3	14.3	10.1
1.00	20	19.7	19.8	15.6	10.3	18.3	15.5	10.3
1.50	20	27.8	19.8	17.0	11.2	18.3	16.8	11.2
2.00	20	35.9	19.8	17.8	12.4	18.3	17.5	12.4

Table 2.15 displays the values for expected sample size for the fixed sample, IP, and GS with and without futility stopping designs. For the GS designs, expected sample sizes are calculated assuming the null hypothesis ($\theta = 0$), the alternative of interest ($\theta = \theta_1$), and assuming an effect of twice the alternative of interest ($\theta = 2\theta_1$).

The results show similar trends in sample sizes as those found in Example 2.1. The fixed sample design will always have the sample size of n_0 , in this case $n_0 = 20$. Expected sample size for the GS designs is always less than the fixed sample size due to possible early stopping and no possible increase in sample size. When no early futility stopping is allowed, the GS expected sample size values will be very close to the fixed study sample size when $\theta = 0$. Otherwise, it shows the biggest sample size benefits for lower variances and larger effect sizes, which both increase the chance of early stopping. Additionally, the GS with futility stopping shows the added benefit of reduced expected sample size under the null hypothesis.

Table 2.16: $E(N_w)$ for Example 2.2: IPIA

γ	IPIA (no futility)			IPIA (futility)		
	$\theta = 0$	$\theta = \theta_1$	$\theta = 2\theta_1$	$\theta = 0$	$\theta = \theta_1$	$\theta = 2\theta_1$
0.50	11.3	10.6	10.0	11.0	10.6	10.0
0.75	14.2	12.8	10.3	13.4	12.8	10.3
1.00	17.8	16.1	11.7	16.4	16.1	11.7
1.50	25.7	24.1	17.9	22.9	23.7	17.9
2.00	33.8	32.5	26.6	29.7	31.6	26.6

Table 2.16 displays the values for expected sample for the IPIA with and without futility stopping designs. The results show similar trends in sample sizes for the IPIA designs as those found in Example 2.1. The IPIA expected samples sizes are lower than the fixed sample design size for low true variance and increase with the true variance as the first stage sample size re-estimation requires larger second stage sample sizes on average. Although the expected sample size rises with γ for the IPIA designs, it still has smaller sample sizes than the IP design (Table 2.15). This is caused by the offsets of early decision sample size benefits and the power protective sample size re-estimation procedure, both of which are characteristics of the IPIA design. Additionally, the IPIA design with futility stopping once again shows the added benefit of reduced sample size under the null hypothesis. The sample size advantage of the IPIA design versus the IP design is extremely evident for an effect of twice the alternative of interest, ($\theta = 2\theta_1$). The IPIA designs offer significant expected sample size reduction for this case due the large chance of early efficacy stopping in the first stage. The effect diminishes as increasing variance calls for more sample size in the second stage and hence decreases first stage power.

By comparing the expected sample sizes from Examples 2.1 and 2.2, one can examine the possible effect of fixed sample study size on expected sample size for these designs. In a proportional sense, there does not seem to be a difference in sample size benefits between the two studies. However, if the cost per subject in the study designs was equal between the larger and smaller studies, than the sample size reductions and increases would be more significant in the larger study (Example 2.1).

2.6 DISCUSSION

In this chapter, I have presented a proposed framework for an internal pilot with interim analysis model for univariate Gaussian linear model hypotheses with fixed predictors. The framework generalizes traditional internal pilots by allowing for early stopping rules at the interim analysis to go along with variance estimate based sample size re-estimation. I have

derived the exact theory needed for the IPIA framework for single degree of freedom GLM hypotheses, including one and two group t -tests with unknown, common variances and other tests of interest. The exact results allow for numerical calculation of type I error rate, power, and expected sample size for various study designs without the need for time-consuming simulations. Many prospective research studies and even clinical trials are not large enough for asymptotic properties to hold. Since the theory in this chapter is not derived using asymptotic results, it will be accurate and valuable for planning smaller studies.

Examples 2.1 and 2.2 highlight some of the different characteristics of the IPIA design while comparing them to results from fixed sample, IP, and two-stage GS designs, all of which are special cases of the IPIA design and theory detailed in this chapter. The size difference between Example 2.1 and Example 2.2 displayed the sensitivity of study characteristics such as type I error rate to sample size. The examples here use large sample distributional assumptions for critical value calculation in order to illuminate features of the designs examined. Each design considered has its advantages and disadvantages.

The fixed sample design has the advantages of a known sample size and a controlled type I error rate, but has an unstable power function affected by the unknown true variance, a nuisance parameter. The goal of the GS design is to allow a study to stop early if effect size differs from the preplanned magnitude, and hence, decrease expected sample size from the fixed sample level. It has been shown in the tables above to achieve this goal under all conditions when futility stopping is allowed, and under situations of high effect size when futility stopping is not included. Using standard Gaussian critical values, the GS designs have an inflation in type I error rate due the critical values not accounting for uncertainty of the variance estimate. Finally, GS power is vulnerable to misspecification of variance at planning stage as shown in Table 2.9. This sensitivity is similar to that found in the fixed sample design and is due to the lack of sample size re-estimation for the final stage sample size. The primary goal of the IP design is to protect study power by re-estimating sample

size through interim power analysis without interim data analysis. As Table 2.9 shows, this goal is greatly achieved by the design. The IP design can also have sample size benefits due to possible sample size reduction at the interim stage if the planning variance value was specified higher than the true parameter value. For high true variance values, the design has higher expected sample sizes than the GS and fixed sample designs. The IP design also has inherent type I error rate inflation dependent on the true variance value. This is typically accounted for in smaller sample studies by adjusting the test statistic or critical value. Adjustment is not made here for comparative purposes.

The IPIA designs seek to incorporate the sample size advantage of the GS by allowing for early stopping and the power protective properties of internal pilots through variance estimate based sample size re-estimation at the interim analysis. Table 2.11 shows that the sample size benefits of the GS design are incorporated into the design as the expected sample sizes for many conditions are significantly lower than for the IP design. The sample size benefits are attained by allowing possible early stopping for efficacy and/or futility at the interim analysis. Also, Table 2.9 shows that the IPIA design does, in fact, greatly achieve the power protective properties of the IP design. This characteristic creates a power function robust to variance misspecification at the planning stage, unlike fixed sample and GS designs.

In addition to the advantages of the IP and GS designs, the IPIA design was shown to inherit some of the elements of concern from the designs. The internal pilot design has type I error rate inflation caused by an unbiased variance estimate used in the final test statistic, with more magnitude of inflation found in small studies where exact theory has the most value. The group sequential designs examined also have type I error rate inflation due to the inappropriate use of large sample critical values. Since the IPIA designs examined in this chapter combine characteristics of these two designs, it has elements of both sources of type I error rate inflation: distributional alignment vulnerability as well as potentially biased

variance estimates used in the second stage test statistic. Because of this, careful adjustments that can control the type I error rate while maintaining the design's benefits must be achieved in order for the IPIA procedure to be useful in practice.

The IPIA procedure as outlined in this chapter is purposefully kept general in many regards. For example, it does not specify mandatory methods for selecting critical values, updating sample size, or selecting the interim stage sample size. The theory developed in this chapter as well as the development of software to assist in calculation would allow for the exploration of many different design possible designs. This would not only be valuable for the development of general study guidelines with positive characteristics, but since studies are not alike, extensive exploration for a specific study during planning stages can allow investigators to customize the procedure for their specific needs. Through comparison to simulation, that analytical calculations were shown in Section 2.5.2 to be highly efficient. Thus, using the exact theory would allow for far more efficient study planning such as graphing power and type I error rate functions over various conditions. Also, the efficiency facilitates new research methods using the exact forms as a component. Procedural strategies for the IPIA model utilizing the exact theory introduced in this chapter here is the major theme of Chapter 3 in this dissertation. Examples from this chapter are revisited with implementation of differing possible design strategies with results examined.

CHAPTER 3. PLANNING PROCEDURES FOR AN INTERNAL PILOT WITH INTERIM ANALYSIS DESIGN

SUMMARY

In Chapter 2 of this dissertation, I introduced exact theory to help plan single degree of freedom internal pilot with interim analysis (IPIA) designs for Gaussian linear models. Here, I discuss and evaluate procedures for planning the studies described in Chapter 2. The goal is to achieve sound study design strategies that control the type I error rate while best maintaining the power and sample size advantages of the IPIA designs. The exact theory allows for simple procedure comparisons and facilitates the development of new procedures.

3.1 INTRODUCTION

3.1.1 Motivation

In Chapter 2, I examined the properties of the proposed two-stage internal pilot with interim analysis (IPIA) procedure and how it compare to properties of special cases such as fixed sample, internal pilot (IP), or two-stage group sequential (GS) procedures. I determined that the IPIA design had sample size advantages compared to the IP design and power advantages compared to the two-stage GS and fixed sample designs. Also, using naive, but common approaches to critical value selection, sample size re-estimation method, and interim sample size calculation showed that, especially in small samples, the IPIA design has the potential for type I error rate inflation. Unless controlled, this can offset the IPIA sample size and power advantages. Smart design procedures and strategies are needed to attain the benefits of the exact IPIA theory derived in Chapter 2.

Pre-specification of critical value selection procedures as well as sample size allocation and re-estimation methods are needed in order to calculate power, type I error rate, and

expected sample size during study planning. Critical values for first stage and second stage must be selected to accurately and strategically allocate the type I error rate while minimizing expected sample size and maintaining power. Sample size determination methods also affect study power and expected sample size. I consider methods that allow implementation of an IPIA design while preserving the type I error rate.

3.1.2 Literature Review

The relevant literature review is largely covered in Chapters 1 and 2 of this dissertation. Some additional background specific to this chapter is included in this section.

A large portion of this chapter is devoted to the problem of potential type I error rate inflation in small sample IPIA studies. I focus on two known sources of inflation. The first source is the use of large sample distributional assumptions in standard Gaussian based sample size re-estimation and critical value techniques. The second source of inflation occurs due to a biased variance estimate that is characteristic of internal pilot based sample size re-estimation designs.

The information based approach to clinical trial monitoring with nuisance parameter based sample size adjustment described by Mehta and Tsiatis (2001) or Tsiatis (2006) was reviewed in Chapter 2. Mehta and Tsiatis described the method for use within group sequential designs that allow for early stopping while updating the estimated maximum sample size at each analysis stage as nuisance parameter estimates are updated. Group sequential stopping bounds along with an inflation factor on needed information (and hence needed sample size) due to multiple testing were advocated. They used standardized test statistics with critical boundary determination based on the error spending technique used. The effects on type I error rate of large sample distributional assumptions are portrayed in the examples of Chapter 2 in this dissertation. For the small and moderate sample size designs considered, clear type I error rate inflation was present.

Within group sequential designs for two group comparisons, some alternative critical value selection methods have been proposed and reviewed. One simple approach suggested by Pocock (1977) shown to work quite well is to take the significance level of Gaussian derived critical values and use them along with sample size to calculate corresponding t distributed critical values. Since the t distribution takes into account the sample size used for estimation in the form of degrees of freedom, it better relates to the uncertainty of the variance estimate used in the test statistic. Although the statistics are sequences and hence have a joint relationship, this simple method approximately controls the type I error rate for group sequential designs.

Shao and Feng (2007) described a Monte Carlo method for calculation of critical values in a small sample group sequential studies. Through simulation they show that their method works well at controlling the type I error rate and maintaining power with an expected increase in expected sample size.

The second source of type I error rate inflation present in IPIA designs is from the use of a downwardly biased variance estimate in study stages following the first. This is the same cause of type I error rate inflation found in unadjusted internal pilot studies; see Proschan and Wittes (2000) or Miller (2005) for details. Various methods for controlling the type I error rate within internal pilots have been considered. Some have the downside of only considering a fraction of the available information in order to create an unbiased variance estimate (Stein, 1945; Zucker et al., 1999). Denne and Jennison (1999) proposed a method based on Stein's test that uses all information about the variance, but includes a degree of freedom adjustment to the final test statistic that does not guarantee bounding of type I error rate, but appears to work well on average. Proschan and Wittes (2000) introduced a method that uses an unbiased variance estimate by fixing weights between the IP stage and the second stage portions of the final variance estimate. Coffey and Muller (2001) introduced a *bounding method* which alters the critical value through an alpha-

adjustment so that the maximum type I error rate inflation across true variance is equal to the target rate. Due to its importance to this chapter, further details on this last method are included below.

For internal pilots, the goal of the bounding method described by Coffey and Muller (2001) is to find the nominal α_b used in critical value determination with maximum type I error rate over possible true variances equal to α_t . Although a definitive proof is not available, Coffey and Muller (2001) provided substantial evidence in support of the hypothesis that there is a single maximum type I error rate as a function of the true variance. The α_b value is calculated by using a doubly iterative algorithm that uses exact internal pilot theory for calculations. The outer loop searches over possible α values for the desired α_b . The inner loop takes advantage of the unimodal characteristic of internal pilot type I error rate over true variance to determine the location and magnitude of the maximum type I error rate for a given α (fixed critical value). The procedure is conservative in that the type I error rate for some true variance values may be less than the target rate, but power properties for the method are quite stable and very near to the unadjusted method. Of note is that the published version of the bounding method uses the originally planned target type I error rate, α_t , for sample size re-estimation. An alternative method currently being considered would use the adjusted level for sample size re-estimation as well as for critical value calculation to better align the assumed distributions. Interim results show that this alternative would greatly close the power gap between the bounding method and an unadjusted method, but would use more sample size.

3.2 THE IPIA MODEL AND PROCEDURE

3.2.1 Notation

Notational conventions will be followed as described in Muller and Stewart (2006, Chapter 1). An $r \times 1$ vector (always a column) is written \mathbf{a} , and an $r \times c$ matrix is written $\mathbf{A} = \{a_{j,k}\}$, with transpose \mathbf{A}' . Let $\mathbf{1}_r$ represent an $r \times 1$ vector of 1's and $Dg(\mathbf{x})$

represents a diagonal matrix with (j, j) element x_j . Furthermore, define \mathbf{I}_r as the $r \times r$ identity matrix with $\mathbf{I}_r = \text{Dg}(\mathbf{1}_r)$. The direct (Kronecker) product is defined as

$$\mathbf{A} \otimes \mathbf{B} = \{a_{j,k} \mathbf{B}\}.$$

Detailed information about all random variables discussed in this paper can be found in Johnson et al. (1994, 1995). Let $\mathbf{x} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ indicate that random vector \mathbf{x} ($n \times 1$) has a vector (multivariate) Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

For $\boldsymbol{\Sigma}$ less than full rank, \mathbf{x} has singular vector Gaussian distribution, written as

$\mathbf{x} \sim S\mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Writing $X \sim \chi^2(\nu, \omega)$ indicates that X follows a non-central chi-square distribution, with ν degrees of freedom and noncentrality ω . Likewise, writing

$X \sim F(\nu_1, \nu_2, \omega)$ indicates that X follows a noncentral F distribution with numerator degrees of freedom ν_1 , denominator degrees of freedom ν_2 , and noncentrality ω . Writing

$\chi^2(\nu)$ or $F(\nu_1, \nu_2)$ implies $\omega = 0$. More generally, writing $X \sim \chi_T^2(\nu; t_L, t_U)$ indicates that X follows a doubly-truncated central chi-square distribution with ν degrees of freedom,

truncated to the interval $[t_L, t_U]$ (Coffey and Muller, 2000). For random variable U with parameters $\gamma_1 \dots \gamma_k$, indicate the cumulative distribution function (CDF) taken at u as

$F_U(u; \gamma_1 \dots \gamma_k)$. As a special case, let $\Phi(z)$ indicate the CDF for the Gaussian(0,1)

distribution, taken at z . Also let $F_U^{-1}(\alpha; \gamma_1 \dots \gamma_k)$ indicate the α quantile of a random variable U with parameters $\gamma_1 \dots \gamma_k$.

3.2.2 The IPIA Model

The internal pilot with interim analysis (IPIA) model discussed in this paper is introduced in Chapter 2 of this dissertation. It can be viewed as a generalization of the two-stage internal pilot model in the GLUM framework as introduced by Coffey and Muller (1999), which includes the t -test as a special case. Let random variable N_w be the final sample size used for the study. Then $N_w = n_1 + N_2 \cdot \mathcal{I}(\text{continue})$ with \mathcal{I} an event indicator equal to 1 if a study is continued at the first stage. So

$$N_w = \begin{cases} n_1 & \text{if study stopped after first stage} \\ N_+ & \text{otherwise} \end{cases} . \quad (3.75)$$

The design leads to interest in two different but intimately connected models. The combined model for the final analysis may be written as

$$\begin{matrix} \mathbf{y}_+ \\ N_+ \times 1 \end{matrix} = \begin{matrix} \mathbf{X}_+ \boldsymbol{\beta} \\ N_+ \times q \times 1 \end{matrix} + \begin{matrix} \mathbf{e}_+ \\ N_+ \times 1 \end{matrix}, \quad (3.76)$$

or

$$\begin{bmatrix} \mathbf{y}_1 \\ n_1 \times 1 \\ \mathbf{y}_2 \\ N_2 \times 1 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ n_1 \times q \\ \mathbf{X}_2 \\ N_2 \times q \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{e}_1 \\ n_1 \times 1 \\ \mathbf{e}_2 \\ N_2 \times 1 \end{bmatrix}, \quad (3.77)$$

with partitioning corresponding to the n_1 and, random, N_2 observations in the first and second samples, respectively. The second sample of size $N_2 = N_+ - n_1$ shown above is only taken if study continuation is determined from the first sample. Also, the special case of $N_+ = n_1$ will cause the full model to collapse to the interim model. Model components include random observed \mathbf{y}_+ ($N_+ \times 1$) (independent sampling units as rows), design matrix of fixed form \mathbf{X}_+ , and unobserved \mathbf{e}_+ such that $\mathbf{e}_+ \sim \mathcal{N}_{N_+}(\mathbf{0}, \sigma^2 \mathbf{I}_{N_+})$. For computational convenience, random values of total sample size, $N_+ = n_1 + N_2$, increase only in multiples of a replication factor, m . For example, a balanced 2-group study design would have $m = 2$. For some \mathbf{X}_0 ($m \times q$), I assume $\mathbf{X}_1 = \mathbf{1}_{k_1} \otimes \mathbf{X}_0$ and $\mathbf{X}_2 = \mathbf{1}_{K_2} \otimes \mathbf{X}_0$, with k_1 and K_2 the number of replications in the first and second samples, respectively. Consequently, the columns of \mathbf{X}_1 and \mathbf{X}_2 span the same space (when $K_2 > 0$) and hence define $r = \text{rank}(\mathbf{X}_1) = \text{rank}(\mathbf{X}_2) = \text{rank}(\mathbf{X}_+)$. In order to simplify computations and some discussions, attention will usually be restricted to a full rank design, that is $\text{rank}(X_0) = q$. The principles of linearly equivalent models allow the restriction without meaningful loss of generality.

The test of interest is $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, with \mathbf{C} a fixed $a \times q$ contrast matrix and $\boldsymbol{\theta} = \mathbf{C}\boldsymbol{\beta}$. Without loss of generality one can assume $\boldsymbol{\theta}_0 = \mathbf{0}$ (see Lemma A.1). For a ‘scientifically

important' effect of interest ($\theta = \theta_1$), a desirable design ensures a target type I error rate (α_t) with sample size appropriate to achieve target power (P_t).

Section 2.2.2 of this dissertation includes tables containing definitions and descriptions of model elements and can be referenced for additional IPIA model details.

3.2.3 The General Procedure

Table 3.1: *General procedure*

Step 1a : Specify α_t , P_t , \mathbf{X}_0 , hypotheses, θ_1 , and σ_0^2
1b : Solve for first stage sample size (n_1)
Step 2 : Collect first n_1 observations
Step 3 : Solve for $N_+ = n_+$, critical values $f_l(n_+)$, $f_u(n_+)$, and $f_+(n_+)$, and F_1
Step 4 : Decide:
If $F_1 < f_l$ then STOP, ACCEPT H_0
If $F_1 \geq f_u$ then STOP, REJECT H_0
If $f_l \leq F_1 < f_u$ then take $n_2 = n_+ - n_1$ additional observations
Step 5 : Solve for F_+
Step 6 : Decide:
If $F_+ < f_+$ then ACCEPT H_0
If $F_+ \geq f_+$ then REJECT H_0

Table 3.1 gives the general procedure for the IPIA model. General steps include selection of n_1 (Step 1b), n_+ (Step 3), and critical values (Step 3). These design factors will be examined in the sections to follow.

3.3 CRITICAL VALUE SELECTION

3.3.1 Overview

Calculation of the three critical values for the study, $f_l(n_+)$, $f_u(n_+)$, and $f_+(n_+)$, must be done following rules pre-specified in the study protocol. The general IPIA procedure described in Section 3.2.3 does not specify a method for determining critical values. Hence, there are countless possible methods from which they can be determined. Ideally, they should be chosen in a way that controls the type I error rate while having good power and expected sample size properties. The critical values may depend on n_+ , the realized value of

N_+ ; however, when it is clear, they will be referred to as f_l , f_u , and f_+ . In this chapter, I focus on the selection of efficacy values (f_u and f_+) and leave discussion for selection of futility bounds to future research.

Efficacy bounds for both stages could be calculated in many ways. The simplest and most naive method would be to assume a large sample distribution (such as Z or χ^2) and ignore the multiple testing issue by using a nominal type I error rate of α_t at each stage. Smarter methods would account for the uncertainty in the variance estimate and/or adjust the nominal type I error rates used for critical value calculation to create more sound designs.

The goal is to choose a limited number of sensible and/or common methods that can be easily employed in practice and compare their study characteristics by way of examples. I plan to adapt the examined methods from the common error rate spending methods of group sequential designs. Once N_+ is determined at the interim stage, one can determine the first stage sample fraction $T = n_1/N_+$ ($0 < T \leq 1$). Based on a specified error-spending method, the value of T can be used to determine what nominal type I error rate to use at the interim and final stages in order to maintain an unbiased hypothesis test. The chosen nominal rates are used to determine the critical values for the two stages. For example, if the error rate spending function specifies that $\alpha_1 = 0.01$ and $\alpha_+ = 0.04$ where α_1 and α_+ are the nominal type I error rates for the interim and final stages, then the large sample critical values could be calculated as $f_u = F_{\chi^2}^{-1}(0.99; 1)$ and $f_+ = F_{\chi^2}^{-1}(0.96; 1)$. For greater values of T (more information), more study error will typically be spent at the first stage.

3.3.2 Distributional Assumptions

For large sample study designs, the uncertainty in the variance estimates used in test statistics becomes minimal. Because of this, variance estimates are often be treated as known quantities during analysis and asymptotically correct distributions (such as Z or χ^2) are commonly used for critical value calculation.

For smaller studies, however, these distributions become increasingly inaccurate and can cause a decrease in study integrity due to type I error rate inflation. A simple adjustment with beneficial results is the use of more accurate distributions that account for the uncertainty of variance estimation with degree of freedom adjustments for sample size (such as t or F). To do this, you take the significance level of Gaussian derived critical values and use them along with sample size to calculate corresponding small sample critical values. For example, if the error rate spending function specifies that $\alpha_1 = 0.01$ and $\alpha_+ = 0.04$ where α_1 and α_+ are the nominal type I error rates for the interim and final stages, then the critical values could be calculated as $f_u = F_F^{-1}(0.99; 1, n_1 - 2)$ and $f_+ = F_F^{-1}(0.96; 1, n_+ - 2)$. In group sequential designs, these distributions have been shown as useful for helping to control the type I error rate of a study (Pocock, 1977). This method could be valuable in smaller IPIA studies by controlling the type I error rate and maintaining the benefits of the IPIA design. In the examples in this chapter, I will compare results using both the large sample and small sample methods of critical value calculation for the IPIA model.

3.3.3 The IPIA Bounding Method

An important characteristic of internal pilot based sample size re-estimation is the potential type I error rate inflation caused by a downward biased variance estimate used in the test statistic calculations (Proschan and Wittes, 2000; Miller, 2005). The effect is more pronounced for smaller studies and should be accounted for during critical value calculation planning. Group sequential based error spending functions, even when more exact theory such as the t distribution are used, do not take this sample size re-estimation based bias into account and hence, may alone be inadequate for the IPIA design.

A proposed critical value selection method for the IPIA design is the *IPIA bounding method*. The method is an adaptation of the design introduced for use in internal pilots by Coffey and Muller (2001) and discussed in Section 3.1.2. The IPIA bounding method employs a similar approach as the IP bounding method. The goal is to predetermine a

nominal target type I error rate, α_b ($\leq \alpha_t$), that when used during the study in place of α_t , controls the true type I error rate to at or below the target level.

Since the IPIA design employs an internal pilot based sample size re-estimation technique, the type I error rate as a function of true variance behaves similarly. That is, holding other study characteristics constant, the type I error rate as a function of true variance is a unimodal curve. Although this is not proven conclusively, I strongly believe it to be true based on all examples considered to date. This characteristic allows for a doubly iterative search algorithm to locate the desired value for nominal target type I error rate, α_b , that can then be used in place of the true target type I error rate, α_t , during study analysis. The outer procedure searches for α_b that bounds the maximum unconditional type I error rate at the target level while the inner procedure determines the location and magnitude of the maximum type I error over variance values. The α_b is then used instead of α_t within the error spending functions that determine the nominal α used for critical value determination. Additionally, for the IPIA bounding method, I will use the new α_b value as the target type I error rate for sample size re-estimation. I have found that this best maintains the power of the IPIA design with only a small cost in sample size.

3.4 SAMPLE SIZE SELECTION METHODS

3.4.1 Sample Size Re-estimation

The sample size re-estimation rule will determine the distribution of N_+ and so is an important consideration in the design affecting power and expected sample size. Section 2.3 describes a procedure for determining the distribution of the total sample size for the IPIA design by taking advantage of the monotone relationship between discrete N_+ and continuous $\hat{\sigma}_1^2$. By calculating the scaled noncentrality parameters needed to achieve a given power, the procedure calculates cut off points into which $\hat{\sigma}_1^2$ must fall in order for a specific value of N_+ to occur. Since the distribution of $\hat{\sigma}_1^2$ is known, the cut-off points give us

$\Pr\{N_+ = n_+\}$ for all possible n_+ allowed in the study design. Some flexibility is allowed within this framework as described below.

In order to determine the needed noncentrality for a given sample size, a critical value must be specified. One common way to do this is to assume an uncorrected α level for the final statistic for sample size estimation purposes. This is common practice in group sequential planning. The resulting sample size is then sometimes multiplied by a design dependent sample size inflation factor in order to take into account the multiple testing element of the study. More complicated ways to determine the α levels are possible, such as adapting it dependent on the information fraction at the interim stage by an α spending function. This would better align the sample size needs to correctly power the study and could do away with the need for the sample size inflation factor. While promising, this creates complications with critical values dependent on sample size and sample size dependent on critical values. More research should be done in this area to better align the sample size re-estimation rules with the distributional realities of the test statistics employed.

Another flexibility in the sample size re-estimation procedure is with the distributional assumptions used. For example, sample size can be calculated assuming either large sample distributions (Z or χ^2) or small sample distributions (t or F) for the distribution of the final test statistic. This affects the noncentrality parameter and hence the sample size distribution.

A disconnect between the distributional assumption used in sample size re-estimation and that employed for critical value determination could cause undesired study properties. For example, using a t distribution during sample size re-estimation would call for sample sizes greater than if a Gaussian distribution is used and creates an over powered study if Gaussian based critical values are employed during analysis. Alternatively, using Gaussian based sample size re-estimation with t distribution based critical values calls for too little sample size creating an underpowered study. The logical method here is to align the distribution type used in sample size re-estimation with that used to calculate critical values

during analysis. Since this kind of alignment was found to work best, it is the process followed for the examples in section 3.5.

3.4.2 Interim Sample Size Selection

The value of the internal pilot sample size, n_1 , should be made at the design stage of the study. The choice is important since lower values give more uncertain estimates of σ^2 while higher values may have better power properties, but can lose sample size benefits. For the IPIA design, the choice of n_1 is completely general and the theory holds for any selection method employed.

For traditional internal pilot designs, Birkett and Day (1994) recommended at least 20 degrees of freedom in the first stage sample; however, Coffey and Muller (1999) showed that this scenario can still produce type I error rate inflation depending on the design and noted the importance of examining properties of specific study designs. A default in literature seems to be taking a designated fraction of the sample size from fixed sample equations using the best guess for the variance at the planning stage, i.e., $\pi \cdot n_0$ for $0 < \pi \leq 1$ and n_0 determined from σ_0^2 . A default choice for π seems to be $1/2$; that is, the size of the first sample is half of the fixed sample study sample size based on σ_0^2 .

With the added element of early stopping, the choice of n_1 for the IPIA model is a more complex matter than in the IP design. Depending on the critical value method employed and true effect size and variance values, a high value of n_1 could possibly have sample size savings due to the changing power at the first stage test statistic. Because of the complex and interactive nature of this factor, I stress that it should always be examined as a study design element keeping in mind the goals of a particular study.

In Section 3.5.6, I numerically examine an example employing three methods for selecting the internal pilot sample size. In Chapter 2, the examples used $n_1 = \pi n_0$ for the value $\pi = 0.5$. The examples in this chapter also use $\pi = 0.5$ and compare its use to the results using $\pi = 0.25$ and 0.75 . The resulting complications will be discussed.

3.5 EXAMPLES

3.5.1 Example Motivation

The results of the examples in this chapter are calculated using the exact theory developed in Chapter 2. The goal is to compare possible study designs within the IPIA framework and find an easy to employ design method that controls the type I error rate while maintaining the power and sample size benefits of the IPIA design. Futility bounds will not be used in this chapter. Their clear benefits were shown in Chapter 2, and refinement of their use is saved for future research.

3.5.2 Example Methods

All numeric calculations in this chapter are done for $\gamma \in \{0.5, 0.75, 1, 1.5, 2\}$ where $\gamma = \sigma^2/\sigma_0^2$. Each example begins by comparing the type I error rates of nine possible designs. Then, selecting study types with superior control of type I error rates, power properties are examined. Finally, expected sample sizes for the designs with both attractive type I error rate and power properties are calculated and described. Since the expected sample sizes of procedures with early stopping depend on the true effect size θ , they will be calculated under three conditions: under the null ($\theta = 0$), under the alternative of interest ($\theta = \theta_1$), and under a true effect twice that of the effect size of interest ($\theta = 2\theta_1$).

Critical values will be calculated by either using the standard Gaussian distribution or by using the t distribution. For procedures with two testing stages, I will employ an O'Brien-Fleming type alpha spending technique using the SAS SEQSCALE function (SAS Institute, 2004). Sample size re-estimation will use α_t to calculate the sample size determining critical values for all values of N_+ . The exception to this is for the bounding methods, which will use the pre-determined adjusted α_b values for sample size determination. The assumed distribution used in sample size re-estimation will mirror the distribution used for analysis critical value determination. All of the designs considered in Sections 3.5.4 and 3.5.5 use

$n_1 = \pi n_0$ for $\pi = 0.5$. The effects of different values for π will be examined in Section 3.5.6.

The nine methods to be compared in Sections 3.5.4 and 3.5.5 are as follows:

- Fixed sample design using standard Gaussian distribution for critical value calculation
- Fixed sample design using t distribution for critical value calculation
- Two-stage group sequential using O'Brien-Fleming critical value calculation based on the standard Gaussian distribution
- Two-stage group sequential using O'Brien-Fleming critical value calculation based on the t distribution
- Internal pilot allowing sample size to decrease from planning estimate ($n_{+,max} = n_1$) using t distribution for sample size re-estimation and critical value calculation
- IPIA design using sample size re-estimation and O'Brien-Fleming critical value calculation based on the standard Gaussian distribution
- IPIA design using sample size re-estimation and O'Brien-Fleming critical value calculation based on the t distribution
- IPIA bounding method* using sample size re-estimation and O'Brien-Fleming critical value calculation based on the standard Gaussian distribution, both using α_b
- IPIA bounding method* using sample size re-estimation and O'Brien-Fleming critical value calculation based on the t distribution, both using α_b

* The bounding method employed is an approximation and does not perfectly bound the type I error rate in all cases. For computational convenience, the prototype software created and used for this research to calculate the adjusted α levels is using a modified version of the bounding method that approximates the algorithm described in Section 3.3. Instead of recalculating the location of maximum type I error rate for each α considered, it does so only for $\alpha = \alpha_t$ (say the location is γ_t). It then finds α_b such that the type I error rate at $\gamma = \gamma_t$ is equal to α_t . In reality the location of the peak can shift slightly with α values which causes the computationally expedient, but approximate IPIA bounding method.

3.5.3 Computational Methods

All programs for the examples were written in SAS/IML (SAS Institute, 2004). Most of the computation for the examples utilizes the exact theory developed in this chapter. The exceptions are the fixed sample and internal pilot designs. The fixed results could be directly calculated using standard distribution functions. The internal pilot calculations were easily obtained utilizing exact internal pilot theory from the freely available GLUMIP 2.0 (Kairalla

et al., 2007) software package. All other results came from use of the exact theory, including the two-stage group sequential designs, which are a special case.

To avoid numerical instability of the calculated integrals, computation was performed using quantile transformations (Glueck and Muller, 2001) of the distributions derived in Section 2.4. For illustration, let $p = F_{\chi^2}(t; \nu)$. Then the integral transformation using $t = F_{\chi^2}^{-1}(p; \nu)$ and $dp = f_{\chi^2}(t; \nu)dt$ allows for integration using finite bounds with better behaved integrands.

3.5.4 Example 3.1 Results

The design parameters for Example 3.1 are summarized in Table 3.2. The nominal target type I error rates used in the bounding methods are α_{bZ} and α_{bt} for the methods using the Z and t distributions, respectively.

Table 3.2: *Design parameters for Example 3.1*

α_t	α_{bZ}	α_{bt}	P_t	θ_1	σ_0^2	n_0	n_1
0.050	0.042	0.047	0.9	1	2	86	44

Example 3.1, a study design of moderate size, was considered by Wittes and Brittain (1990) and Coffey and Muller (1999) in an internal pilot framework and previously in Chapter 2 for the IPIA.

Table 3.3: *Type I error rates $\times 100$ for Example 3.1*

γ	Fixed		GS		IP	IPIA		Bounding	
	Z	t	Z	t	t	Z	t	Z	t
0.50	5.3	5.0	5.5	5.0	5.2	5.7	5.1	4.9	4.8
0.75	5.3	5.0	5.5	5.0	5.4	6.0	5.3	5.1	5.0
1.00	5.3	5.0	5.5	5.0	5.3	5.9	5.4	5.0	5.0
1.50	5.3	5.0	5.5	5.0	5.2	5.6	5.3	4.7	5.0
2.00	5.3	5.0	5.5	5.0	5.2	5.4	5.2	4.5	4.8

Table 3.3 displays the type I error rate values for each of the nine designs described in Section 3.5.2. To see results as a function of variance, values are calculated for true variances corresponding to values of $\gamma = \sigma^2/\sigma_0^2$ with $\gamma \in \{0.5, 0.75, 1, 1.5, 2\}$.

It is clear from the table that for this moderately sized study design, type I error rate inflation is a concern, especially for the designs using the standard Gaussian distribution for sample size re-estimation and critical value calculation. The effect of not accounting for the uncertainty in the variance estimate becomes extremely clear in the fixed sample and GS designs, which control the type I error rate when the t distribution is used. Within the IPIA design, the use of the t distribution significantly decreases the type I error rate inflation, leaving the internal pilot based inflation caused by the biased variance estimate. The IPIA bounding methods control the type I error rate quite well in both cases, however, the type I error rate is more stable across γ for the t distribution based design.

From the results in Table 3.3, I claim that for moderately sized study designs, the use of distributions taking into account the uncertainty of the variance estimate are preferred to the use large sample calculations which ignore the uncertainty. The focus of the examination of power will be on the five designs using the t distribution for this example.

Table 3.4: *Power $\times 100$ for Example 3.1*

γ	Fixed(t)	GS(t)	IP(t)	IPIA(t)	Bounding(t)
0.50	99.6	99.3	92.9	92.6	92.4
0.75	96.3	96.0	90.6	90.1	90.1
1.00	90.0	89.6	90.0	89.7	89.7
1.50	75.4	74.8	89.6	89.2	89.2
2.00	63.0	62.4	89.3	89.0	89.0

Table 3.4 displays the values for unconditional power for each of the five designs that use the t distribution for sample size re-estimation and critical value calculation.

The table makes a clear distinction between the designs with and without sample size re-estimation abilities. All of the designs basically achieve the target power when the planning assumption for variance is true ($\gamma = 1$). However, if the true value of γ varies either lower or higher, the fixed sample and group sequential methods become either over or under powered. All three of the methods with sample size re-estimation have very similar and stable power properties over the true variance values considered. Since this power

protective behavior is desirable, I will continue by looking at sample size properties for the sample size re-estimation procedures.

Table 3.5: $E(N_w)$ for Example 3.1

γ	IP(t)	IPIA(t)			Bounding(t)		
		$\theta = 0$	$\theta = \theta_1$	$\theta = 2\theta_1$	$\theta = 0$	$\theta = \theta_1$	$\theta = 2\theta_1$
0.50	48.1	48.1	45.2	44.0	48.6	45.4	44.0
0.75	66.2	66.0	56.4	44.3	67.1	57.5	44.4
1.00	87.0	86.8	76.9	48.7	88.3	78.8	49.4
1.50	129.0	128.9	123.9	87.7	131.1	126.5	91.4
2.00	171.0	171.0	168.9	146.7	173.9	172.1	151.5

Table 3.5 displays the values for expected sample size for the IP, IPIA, and IPIA Bounding designs. For the IPIA and IPIA Bounding designs, expected sample sizes are calculated assuming the null hypothesis ($\theta = 0$), the alternative of interest ($\theta = \theta_1$), and assuming a true effect size twice the effect of interest ($\theta = 2\theta_1$).

Since the IPIA designs considered here do not include early futility stopping and there is little chance for early stopping when the null is true, the expected sample sized at $\theta = 0$ comes very close to the IP design. The advantages of the IPIA methods become more apparent here as θ increases. The early stopping capability makes the IPIA method more efficient than the IP by lowering expected sample size and saving resources.

The IPIA bounding method uses a little more sample size on average than its unadjusted counterpart. This is due to the use of the adjusted α level in the sample size re-estimation procedure and first stage critical value calculation. The small increase in sample size, however, is offset by the control of the type I error rate for the design. For this example, the IPIA bounding method is the only one of those considered to control the type I error rate, maintain a stable power, and hold a sample size advantage over the IP design.

3.5.5 Example 3.2 Results

The design parameters for Example 3.2 are summarized in Table 3.6. The nominal target type I error rates used in the bounding methods are α_{bZ} and α_{bt} for the methods using the Z and t distributions, respectively.

Table 3.6: *Design parameters for Example 3.2*

α_t	α_{bZ}	α_{bt}	P_t	θ_1	σ_0^2	n_0	n_1
0.050	0.019	0.038	0.9	1.6	1	20	10

Example 3.2 is a smaller study than Example 3.1 and was considered by Coffey and Muller (1999) as well as in Chapter 2 in order to explore small sample study properties. Values for type I error rate, power, and expected sample size are analytically calculated using the exact theory developed in Chapter 2 under the design conditions described in Section 3.5.2.

Table 3.7: *Type I error rates $\times 100$ for Example 3.2*

γ	Fixed		GS		IP	IPIA		Bounding	
	Z	t	Z	t	t	Z	t	Z	t
0.50	6.6	5.0	7.8	5.1	5.5	8.8	5.4	4.9	4.2
0.75	6.6	5.0	7.8	5.1	6.2	9.3	6.0	5.2	4.7
1.00	6.6	5.0	7.8	5.1	6.5	9.5	6.5	5.2	5.1
1.50	6.6	5.0	7.8	5.1	6.5	9.4	6.6	4.7	5.2
2.00	6.6	5.0	7.8	5.1	6.2	8.8	6.4	4.1	4.9

Table 3.7 displays the values for type I error rate for each of the nine designs described in Section 3.5.2. To see results as a function of variance, values are calculated for true variances corresponding to values of $\gamma = \sigma^2/\sigma_0^2$ with $\gamma \in \{0.5, 0.75, 1, 1.5, 2\}$. The results above show that the potential of type I error rate inflation becomes more dangerous and extensive for small sample studies. The cost of not accounting for the uncertainty in the variance estimate is clear in the fixed sample, GS, and IPIA designs. While the fixed sample and group sequential designs largely control the type I error rate when the t distribution is used instead, the IPIA retains a much decreased, but significant inflation. This is due to the

internal pilot based inflation caused by the biased variance estimate. All of the inflation levels become magnified with the small sample study. The IPIA bounding methods control the type I error rate quite well in both cases, however, the extreme nature of the α correction needed for the Z based design makes it undesirable.

From the results in Table 3.7, I claim that for small sample study designs, the use of distributions taking into account the uncertainty of the variance estimate are greatly preferred to the use large sample calculations, which ignore the uncertainty. The examination of power will focus on the five designs using the t distribution for this example.

Table 3.8: $Power \times 100$ for Example 3.2

	Fixed(t)	GS(t)	IP(t)	IPIA(t)	Bounding(t)
0.50	99.8	99.6	92.9	95.8	95.4
0.75	97.4	97.1	90.6	92.9	92.5
1.00	92.2	91.9	90.0	91.0	90.6
1.50	78.9	78.3	89.6	88.7	88.2
2.00	66.8	66.2	89.3	87.3	86.8

Table 3.8 displays the values for unconditional power for each of the five designs that use the t distribution for sample size re-estimation and critical value calculation.

The results for power closely mirror the results obtained for Example 3.1. The methods with sample size re-estimation all have superior power properties over the true variance values considered. Since sample size re-estimation procedures in this example have beneficial power protective behavior, I will examine their sample size properties.

Table 3.9: $E(N_w)$ for Example 3.2

γ	IP(t)	IPIA(t)			Bounding(t)		
		$\theta = 0$	$\theta = \theta_1$	$\theta = 2\theta_1$	$\theta = 0$	$\theta = \theta_1$	$\theta = 2\theta_1$
0.50	12.3	12.3	11.5	10.2	13.0	12.2	10.4
0.75	15.9	15.8	14.9	11.8	16.9	16.0	12.8
1.00	19.7	19.7	18.8	15.2	21.1	20.4	17.0
1.50	27.8	27.7	27.1	24.2	29.7	29.3	26.9
2.00	35.9	35.9	35.5	33.5	38.5	38.2	36.7

Table 3.9 displays the values for expected sample size calculated for the IP, IPIA, and IPIA Bounding designs. The values are calculated for values of $\gamma = \sigma^2/\sigma_0^2$ with $\gamma \in \{0.5, 0.75, 1, 1.5, 2\}$. For the IPIA and IPIA Bounding designs, expected sample sizes are calculated assuming the null hypothesis ($\theta = 0$), the alternative of interest ($\theta = \theta_1$), and assuming a true effect size twice the effect of interest ($\theta = 2\theta_1$).

The sample size trends for this small sample study are similar to those found in Example 3.1. For this small sample example, the IPIA bounding method is the only method of those considered that controls the type I error rate, maintains a stable power function, and it largely holds a sample size advantage over the IP design. The IPIA bounding method thus shows it largely achieves its goals even in the small sample design considered here.

3.5.6 Interim Sample Size Selection Results

Example 3.1 using the IPIA bounding method based on the t distribution is here further examined in order to describe the effects of the choice of interim sample size. Study designs using each case of $\pi \in \{0.25, 0.5, 0.75\}$ are used to determine the initial sample size. The middle value ($\pi = 0.5$) corresponds to the case described in Section 3.5.4.

Table 3.10: *Type I error rates $\times 100$ for $\pi \in \{0.25, 0.5, 0.75\}$*

	$\alpha_b = 0.043$	$\alpha_b = 0.047$	$\alpha_b = 0.048$
γ	$n_1 = 22$	$n_1 = 44$	$n_1 = 66$
0.50	5.0	4.8	4.8
0.75	4.8	5.0	4.8
1.00	4.7	5.0	5.0
1.50	4.5	5.0	5.0
2.00	4.3	4.8	4.9

Table 3.10 displays the values for type I error rate calculated using the exact theory described in Chapter 2 for the t distribution based IPIA bounding method described in Section 3.3.3. Values are calculated for true variances corresponding to values of $\gamma = \sigma^2/\sigma_0^2$ with $\gamma \in \{0.5, 0.75, 1, 1.5, 2\}$ as well as for $\pi \in \{0.25, 0.5, 0.75\}$. The values for α_b were computed separately for each of the three designs.

The results for the bounding method are shown to successfully control the type I error rate at the target level for each of the designs. Also, the location and magnitude of type I error rate peak varies depending on the size of the interim stage. The study design with smaller interim sample size has a type I error peak at a lower variance and has a steeper type I error rate curve over true variance than the other designs. The type I error rate peak shifts higher and the curve becomes flatter as the interim sample size increases.

Table 3.11: *Power $\times 100$ for $\pi \in \{0.25, 0.5, 0.75\}$*

γ	$n_1 = 22$	$n_1 = 44$	$n_1 = 66$
0.50	89.7	92.4	97.8
0.75	88.7	90.1	92.0
1.00	88.1	89.7	90.0
1.50	87.5	89.2	89.6
2.00	87.2	89.0	89.5

Table 3.11 displays the values for power calculated using the exact theory described in Chapter 2 for the t distribution based IPIA bounding method.

The three design cases considered exhibit differences in power properties over true variance values. The larger interim sample size ($n_1 = 66$) causes the study to be over powered for small γ . This is due to the study already having more than enough sample size without the ability to decrease. Conversely, the study with small interim sample size ($n_1 = 22$) has trouble achieving the power of the other designs. This is most likely due to the inaccuracy of the variance estimate used for sample size re-estimation purposes at the interim stage. The power properties for all designs considered here are very good compared to non sample size re-estimation procedures, however, the larger and smaller studies seem to tend to be more over and under powered over various conditions.

Table 3.12: $E(N_w)$ for $\pi \in \{0.25, 0.5, 0.75\}$

γ	$n_1 = 22$			$n_1 = 44$			$n_1 = 66$		
	$\theta = 0$	$\theta = \theta_1$	$\theta = 2\theta_1$	$\theta = 0$	$\theta = \theta_1$	$\theta = 2\theta_1$	$\theta = 0$	$\theta = \theta_1$	$\theta = 2\theta_1$
0.50	46.8	43.6	30.0	48.6	45.4	44.0	66.0	66.0	66.0
0.75	68.6	67.1	55.8	67.1	57.5	44.4	76.8	67.2	66.0
1.00	90.5	89.8	83.6	88.3	78.8	49.4	88.1	75.5	66.0
1.50	134.2	134.1	132.5	131.1	126.5	91.4	130.3	114.2	71.1
2.00	178.0	177.9	177.5	173.9	172.1	151.5	173.0	162.7	100.9

Table 3.12 displays the values for expected sample size calculated using the exact theory described in Chapter 2 for the t distribution based IPIA bounding method. Expected sample sizes are calculated assuming the null hypothesis ($\theta = 0$), the alternative of interest ($\theta = \theta_1$), and assuming a true effect size twice the effect of interest ($\theta = 2\theta_1$).

The effect of interim sample size on expected sample size is fairly complex. Under the null ($\theta = 0$), expected sample size is higher for low γ in the $\pi = 0.75$ and similar for high γ . Under the alternative of interest ($\theta = \theta_1$), the middle sized study ($\pi = 0.5$) saves sample size for lower γ and the larger study ($\pi = 0.75$) saves sample size for high γ . For a large true effect size ($\theta = \theta_1$), the small study saves sample size for low true variance ($\gamma = 0.5$), the middle sized study works best for middle variances (like $\gamma = 1.0$), and the larger study works best when variance is large.

The process behind the effects shown in the tables above is complicated since it involves a combination of minimum sample size, precision of estimates, and power at interim stage. All of these factors are sensitive to the interim stage size, the true effect size, and the true variance.

3.6 DISCUSSION

In order to conduct an IPIA study, design features such as critical value selection procedures as well as sample size allocation and re-estimation methods must be pre-specified. In this chapter, I have examined the use of easy to employ study design characteristics within the IPIA framework using the exact theory introduced in Chapter 2.

The exact theory allowed for fast and accurate calculations to facilitate the comparison of designs and also allowed us to develop new methods with the theory as a basis. I proposed simple design strategies as well as the new IPIA bounding method and compared their type I error rate, power, and expected sample size characteristics under various conditions through examples. The results suggest that IPIA designs can be very useful in maintaining study integrity while minimizing needed time and resources. Additionally, the bounding method has all of the desirable properties wanted in a two-stage IPIA design.

I compared the use of t distribution with the standard Gaussian distribution based methods of critical value determination. Simply using the t distribution for sample size re-estimation and critical value determination for small to moderate sample study designs can go a long way towards controlling the type I error rate inflation inherent to the design. This works well since it takes into account the uncertainty of the variance estimate due to sample size. While this very simple method works in the examples considered for improving the alignment of the critical values with the test statistics, it is not an exhaustive conclusion as the best possible method since the final test statistic does not exactly follow a t distribution. Incorporating other group sequential critical value selection methods such as that described by Shao and Feng (2007) with the IPIA design could increase study characteristics and should be considered.

In addition to the distribution type assumed for critical value selection, sample size re-estimation methods have another source of potential problems due to a downwardly biased variance estimate. The impact of this characteristic is most apparent in small sample studies, such as those considered here. In order to control for this source of type I error rate inflation, I introduced the IPIA bounding method. The method works within the group sequential bound framework to achieve a test with maximum type I error at the target level. In the examples considered, the conservative test was shown to work quite well for achieving the

goals of the IPIA design. It achieves an unbiased error rate while maintaining the IPIA power and having sample size benefits.

Finally, in this chapter I examined the effects of using different sample sizes for the interim stage of an IPIA design. Most of the examples considered in this chapter and in Chapter 2 used a $n_1 = n_0 \cdot \pi$ for $\pi = 0.5$ technique. However, I conclude that the choice is important and far from clear cut. The example showed that differing sizes have benefits under different conditions. Some of the sample principles apply here as in the internal pilot case such as a larger interim size giving a better variance estimate. However, the interplay between re-estimation and testing at the interim stage causes the choice of interim size to be more complex in the IPIA setting. Because of the sensitivity of type I error, power, and expected sample size to the interim sample size of the study, its effects should be explored in more detail and in particular should be explored during study planning for a particular study. With the study goals in mind, an educated decision can allow researchers to achieve the benefits they most desire from the design.

In addition to the results described here, there are many areas where important future methodological work is needed. For example, the sample size re-estimation method employed during a study affects the distribution of sample size and is thus an important consideration for study design. A common approach is to use the target type I error rate to determine a critical value to base sample size needs on during the interim analysis. This approach makes matters simple, but could possibly be improved upon by considering the amount of error spent during the interim analysis. I believe that aligning the alpha levels would better the alignment of the final test statistic with that assumed at the interim analysis and thus create a more efficient study. Another consideration related to sample size re-estimation technique is the assumed distribution of the final test statistic during the interim power analysis. I conclude that power is best maintained at the target level by aligning the distribution used in sample size calculation by that used for critical value calculation.

Finally, in addition to the efficacy stopping bounds considered in this chapter, important future work should include the incorporation of smart futility stopping bounds for the first stage analysis of the IPIA design. As shown in Chapter 2 of this dissertation, this can save much time and resources when little or no effect is present in a study. The resources saved could then be allocated to other promising studies and thus further important scientific research.

CHAPTER 4. INTERNAL PILOT WITH INTERIM ANALYSIS FOR MULTIPLE DEGREE OF FREEDOM HYPOTHESIS TESTS

SUMMARY

In Chapter 2 of this dissertation, I introduced the procedure and theory for an internal pilot with interim analysis (IPIA) design for use with single degree of freedom hypothesis tests. In this chapter, I will again focus on the IPIA design for Gaussian linear models as introduced in Chapter 2, but for use in more complex designs such as multi-group comparisons. In this case C , the contrast matrix in the General Linear Univariate Model (GLUM) framework, has more than one row ($a > 1$), creating multiple degree of freedom hypotheses. I introduce exact theory that can be used in small sample situations to plan studies with complex hypotheses. The theory includes an exact computable form for the conditional joint distribution of the first and second stage test statistics. Together with the results from Chapter 2, the new results allow calculation of power, type I error rate, and expected sample size for *any* univariate linear model with fixed predictors and Gaussian errors. Examples compare study characteristics with a fixed sample design as well as with internal pilot and two-stage group sequential designs, all of which are special cases within the IPIA framework.

4.1 INTRODUCTION

4.1.1 Motivation

When planning clinical trials and other studies, researchers would like to ensure they have an appropriate sample size to detect an effect of interest for a given target type I error rate and power. Researchers and sponsors would also like to have the ability to reach early

decisions when hypothesis outcomes are clear. In Chapter 2 of this dissertation, I focused on single degree of freedom hypotheses, derived the needed exact theory, and laid out the procedure for a two-stage internal pilot with interim analysis (IPIA). The design maintains power by re-estimating sample size needs and can save resources by allowing for early stopping. While most methodological research in internal pilot theory for continuous data involves only the independent groups t -test setting, not all study designs, or even all clinical trials involve only one or two groups. For example, a recently published study by Totonchi and Guyuron (2007) compared treatments suggested for reducing the postoperative edema and bruising associated with rhinoplasty. A total of 48 rhinoplasty patients were randomized to one of two treatment plans or a control group and a three-group analysis of variance (ANOVA) performed to evaluate results. This is an example of a small sample 3-group design that could benefit from the IPIA setting.

Generalizing the exact distributional theory to more complex hypotheses involving multiple group comparisons with computable formulae for power and sample size would allow researchers to accurately explore properties for such designs, even in small samples, before undertaking a study. The importance of small sample theory is explicitly highlighted within the NIH Roadmap (Clinical and Translational Science Awards, RFA-RM-07-002 U54). The exact theory would allow for efficient study planning without the need for simulations.

4.1.2 Literature Review

The relevant literature review is largely covered in Chapters 1 and 2. Some additional background specific to this chapter is included in this section.

While most theory for continuous Gaussian data in internal pilot and group sequential designs has been focused on one and two group t tests, some effort has been made to generalize the methods to more complex hypothesis tests. Spurrier (1982) presented two-stage group sequential tests of hypothesis in the general linear univariate model with

normally distributed and independent errors. Additionally, Jennison and Turnbull (1991, 1997) described distributional theory for multiple stage group sequential t , χ^2 , and F tests.

In order to accommodate more complex designs for internal pilots, Coffey and Muller (1999, 2000, 2001) extended the idea into any univariate linear model with fixed predictors and Gaussian errors. They derived computable forms for the exact distribution of the test statistic. The exact theory allows for flexibility of hypotheses in combination with accurate and efficient study planning for small sample internal pilot designs.

4.2 IPIA MODEL AND PROPERTIES

4.2.1 Notation

Notational conventions will be followed as described in Muller and Stewart (2006, Chapter 1). An $r \times 1$ vector (always a column) is written \mathbf{a} , and an $r \times c$ matrix is written $\mathbf{A} = \{a_{j,k}\}$, with transpose \mathbf{A}' . For full rank matrix \mathbf{A} , the inverse of the transpose equals the transpose of the inverse, so I use $\mathbf{A}^{-t} = (\mathbf{A}')^{-1} = (\mathbf{A}^{-1})'$. Throughout $\mathbf{1}_r$ represents an $r \times 1$ vector of 1's and $\text{Dg}(\mathbf{x})$ represents a diagonal matrix with (j, j) element x_j . Furthermore, define $\mathbf{I}_r = \text{Dg}(\mathbf{1}_r)$ as the $r \times r$ identity matrix. The direct (Kronecker) product is defined as $\mathbf{A} \otimes \mathbf{B} = \{a_{j,k}\mathbf{B}\}$.

Detailed information about all random variables discussed in this paper can be found in Johnson et al. (1994, 1995) and Muller and Stewart (2006). Writing $\mathbf{x} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ indicates that random vector \mathbf{x} ($n \times 1$) has a vector (multivariate) Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. For $\boldsymbol{\Sigma}$ less than full rank, \mathbf{x} has singular vector (multivariate) Gaussian distribution, written as $\mathbf{x} \sim \mathcal{SN}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Writing $\mathbf{x} \sim (\mathcal{S})\mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ indicates the possibility of a singular distribution. Furthermore, $\mathbf{X} \sim \mathcal{N}_{n,p}(\mathbf{M}, \boldsymbol{\Xi}, \boldsymbol{\Sigma})$ indicates that random matrix \mathbf{X} follows a matrix Gaussian distribution, which implies that $\text{vec}(\mathbf{X}) \sim (\mathcal{S})\mathcal{N}_{np}[\text{vec}(\mathbf{M}), \boldsymbol{\Sigma} \otimes \boldsymbol{\Xi}]$. If $\mathbf{X} \sim \mathcal{N}_{n,p}(\mathbf{M}, \mathbf{I}_n, \boldsymbol{\Sigma})$, then $\mathbf{X}'\mathbf{X} \sim W_p(n, \boldsymbol{\Sigma}, \mathbf{M}'\mathbf{M})$ indicates that $\mathbf{X}'\mathbf{X}$ a noncentral Wishart distribution with n degrees of freedom, *shift* parameter $\boldsymbol{\Delta} = \mathbf{M}'\mathbf{M}$, and *noncentrality* $\boldsymbol{\Omega} = \mathbf{M}'\mathbf{M}\boldsymbol{\Sigma}^{-1}$.

Writing $W_p(n, \Sigma)$ implies $\mathbf{M} = \mathbf{0}$. Muller and Stewart (2006, Chapters 8 and 10) presented more details about the matrix Gaussian and Wishart distributions.

Writing $X \sim \chi^2(\nu, \omega)$ indicates that X follows a non-central chi-square distribution, with ν degrees of freedom and noncentrality ω . Likewise, writing $X \sim F(\nu_1, \nu_2, \omega)$ indicates that X follows a noncentral F distribution with numerator degrees of freedom ν_1 , denominator degrees of freedom ν_2 , and noncentrality ω . Writing $\chi^2(\nu)$ or $F(\nu_1, \nu_2)$ implies $\omega = 0$. More generally, writing $X \sim \chi_T^2(\nu; t_L, t_U)$ indicates that X follows a doubly-truncated central chi-square distribution with ν degrees of freedom, truncated to the interval $[t_L, t_U]$ (Coffey and Muller, 2000).

For random variable U with parameters $\gamma_1 \dots \gamma_k$, indicate the cumulative distribution function (CDF) taken at u as $F_U(u; \gamma_1 \dots \gamma_k)$. As a special case, $\Phi(z)$ indicates the CDF for the standard Gaussian distribution (mean zero, variance one), taken at z . Also $F_U^{-1}(\alpha; \gamma_1 \dots \gamma_k)$ indicates the α quantile of a random variable U with parameters $\gamma_1 \dots \gamma_k$.

4.2.2 The IPIA Model

The internal pilot with interim analysis (IPIA) models discussed in this paper can be viewed as generalizations of the internal pilot model in the GLUM framework as introduced by Coffey and Muller (1999). However, due to the possibility of early stopping, notational changes are necessary. In an IP design, N_+ with $n_{+, \min} \leq N_+ \leq n_{+, \max}$ is the random final sample size that is calculated using $\hat{\sigma}_1^2$, the variance estimate from the interim sample. For the IPIA model, N_+ ($n_1 \leq N_+ \leq n_{+, \max}$) is also a random variable based on $\hat{\sigma}_1^2$. However, due to the possibility of early stopping, it is not necessarily the final sample size for the study. Hence for clarity N_w indicates the random final sample size used for the study. Furthermore $N_w = n_1 + N_2 \cdot \mathcal{I}(\text{continue})$ with \mathcal{I} an event indicator equal to 1 if a study is continued at the first stage. Equivalently

$$N_w = \begin{cases} n_1 & \text{if study stopped after first stage} \\ N_+ & \text{otherwise} \end{cases} . \quad (4.1)$$

The design leads to interest in three different but intimately connected models. The combined model for the final analysis may be written as

$$\begin{matrix} \mathbf{y}_+ \\ N_+ \times 1 \end{matrix} = \begin{matrix} \mathbf{X}_+ \boldsymbol{\beta} \\ N_+ \times q \times 1 \end{matrix} + \begin{matrix} \mathbf{e}_+ \\ N_+ \times 1 \end{matrix}, \quad (4.2)$$

or

$$\begin{bmatrix} \mathbf{y}_1 \\ n_1 \times 1 \\ \mathbf{y}_2 \\ N_2 \times 1 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ n_1 \times q \\ \mathbf{X}_2 \\ N_2 \times q \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{e}_1 \\ n_1 \times 1 \\ \mathbf{e}_2 \\ N_2 \times 1 \end{bmatrix}, \quad (4.3)$$

with partitioning corresponding to the fixed n_1 and random N_2 observations in the first and second samples, respectively. Here the second sample of size $N_2 = N_+ - n_1$ shown above is only taken if study continuation is determined from the first sample. Also, the special case of $N_+ = n_1$ will cause the full model to collapse to the interim model. Model components include random observed \mathbf{y}_+ ($N_+ \times 1$) (independent sampling units as rows), design matrix of fixed form \mathbf{X}_+ , and unobserved \mathbf{e}_+ such that $\mathbf{e}_+ \sim \mathcal{N}_{N_+}(\mathbf{0}, \sigma^2 \mathbf{I}_{N_+})$. For computational convenience, random values of total sample size, $N_+ = n_1 + N_2$, increase only in multiples of a replication factor, m . For example, a balanced 3-group study design would have $m = 3$. For some \mathbf{X}_0 ($m \times q$), I assume $\mathbf{X}_1 = \mathbf{1}_{k_1} \otimes \mathbf{X}_0$ and $\mathbf{X}_2 = \mathbf{1}_{K_2} \otimes \mathbf{X}_0$, with fixed k_1 and random K_2 the number of replications in the first and second samples, respectively.

Consequently, the columns of \mathbf{X}_1 and \mathbf{X}_2 span the same space (when $K_2 > 0$) and hence define $r = \text{rank}(\mathbf{X}_1) = \text{rank}(\mathbf{X}_2) = \text{rank}(\mathbf{X}_+)$. In order to simplify computations and some discussions, attention will usually be restricted to a full rank design, that is $\text{rank}(\mathbf{X}_0) = q$. The principles of linearly equivalent models allow the restriction without meaningful loss of generality.

The test of interest is $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, with $\boldsymbol{\theta} = \mathbf{C}\boldsymbol{\beta}$ and \mathbf{C} a fixed $a \times q$ contrast matrix. The results of this chapter will focus on hypotheses with $a > 1$. Without loss of generality I assume $\boldsymbol{\theta}_0 = \mathbf{0}$ (Lemma A.1). For a ‘scientifically important’ effect of interest ($\boldsymbol{\theta} = \boldsymbol{\theta}_1$), a

preferable design ensures a target type I error rate (α_t) with sample size appropriate to achieve target power (P_t).

Throughout, subscript $s \in \{1, 2, +\}$ indicates a value for either the model based on the internal pilot (first) sample, the second sample, or the total combined sample (conditioned on $N_+ = n_+$). Error degrees of freedom are $\nu_s = n_s - r$.

Section 2.2.2 of this dissertation includes tables containing definitions and descriptions of model elements and can be referenced for additional IPIA model details.

4.2.3 IPIA Properties

The IPIA model properties developed in Section 2.2.3 (equations 2.11-2.55) were developed for the general case of a numerator degrees of freedom hypotheses and hence still hold for hypotheses considered in this chapter ($a > 1$). They are not all included here but will be cited when necessary. Additional model properties are needed especially for $a > 1$ case and are included in this section and in Appendix B for reference.

Defining the following $n_+ \times n_+$ matrices facilitates deriving the distributions needed to calculate IPIA probabilities for the $a > 1$ case.

$$\mathbf{A}_{h+} = \mathbf{X}_+(\mathbf{X}'_+\mathbf{X}_+)^{-1}\mathbf{C}'\mathbf{M}_+^{-1}\mathbf{C}(\mathbf{X}'_+\mathbf{X}_+)^{-1}\mathbf{X}'_+ \quad (4.4)$$

$$\mathbf{A}_{h1} = \begin{bmatrix} \mathbf{X}_1(\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{C}'\mathbf{M}_1^{-1}\mathbf{C}(\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \quad (4.5)$$

$$\mathbf{A}_{h2} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2(\mathbf{X}'_2\mathbf{X}_2)^{-1}\mathbf{C}'\mathbf{M}_2^{-1}\mathbf{C}(\mathbf{X}'_2\mathbf{X}_2)^{-1}\mathbf{X}'_2 \end{bmatrix} \quad (4.6)$$

$$\mathbf{A}_{e+} = \mathbf{I}_{n_+} - \mathbf{H}_+ \quad (4.7)$$

$$\mathbf{A}_{e1} = \begin{bmatrix} \mathbf{I}_{n_1} - \mathbf{H}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0}_{n_2 \times n_2} \end{bmatrix} \quad (4.8)$$

$$\mathbf{A}_{e2} = \begin{bmatrix} \mathbf{0}_{n_1 \times n_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{n_2} - \mathbf{H}_2 \end{bmatrix} \quad (4.9)$$

$$\mathbf{A}_{ep} = \mathbf{A}_{h1} + \mathbf{A}_{h2} - \mathbf{A}_{h+} \quad (4.10)$$

$$\mathbf{A}_{eb} = \mathbf{A}_{e+} - \mathbf{A}_{e1} - \mathbf{A}_{e2} - \mathbf{A}_{ep} \quad (4.11)$$

Additionally, I define the following matrices

$$\mathbf{X}_{1*} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{0} \end{bmatrix}_{n_+ \times q} \quad (4.12)$$

$$\mathbf{X}_{2*} = \begin{bmatrix} \mathbf{0} \\ \mathbf{X}_2 \end{bmatrix}_{n_+ \times q} \quad (4.13)$$

$$\mathbf{B} = (\mathbf{X}'_0 \mathbf{X}_0)^{-1} \mathbf{C}' \mathbf{M}_0^{-1} \mathbf{C} (\mathbf{X}'_0 \mathbf{X}_0)^{-1} . \quad (4.14)$$

Lemma 4.1 defines a relationship that will be useful in subsequent results.

Lemma 4.1 For $s \in \{1, 2, +\}$, the following holds.

$$\mathbf{B} \mathbf{X}'_s \mathbf{X}_s \mathbf{B} = k_s \mathbf{B} . \quad (4.15)$$

A proof for Lemma 4.1 is in Appendix B.

Lemma 4.2 For the model in equation 4.3 interpreted as a fixed n_+ design, the following hold for the matrices defined in equations 4.4-4.11.

a. The matrices are all symmetric.

b. \mathbf{A}_{h+} , \mathbf{A}_{h1} , and \mathbf{A}_{h2} can be re-written as

$$\mathbf{A}_{h+} = k_+^{-1} \mathbf{X}_+ \mathbf{B} \mathbf{X}'_+ = k_+^{-1} \begin{bmatrix} \mathbf{X}_1 \mathbf{B} \mathbf{X}'_1 & \mathbf{X}_1 \mathbf{B} \mathbf{X}'_2 \\ \mathbf{X}_2 \mathbf{B} \mathbf{X}'_1 & \mathbf{X}_2 \mathbf{B} \mathbf{X}'_2 \end{bmatrix} \quad (4.16)$$

$$\mathbf{A}_{h1} = k_1^{-1} \mathbf{X}_{1*} \mathbf{B} \mathbf{X}'_{1*} = k_1^{-1} \begin{bmatrix} \mathbf{X}_1 \mathbf{B} \mathbf{X}'_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \quad (4.17)$$

$$\mathbf{A}_{h2} = k_2^{-1} \mathbf{X}_{2*} \mathbf{B} \mathbf{X}'_{2*} = k_2^{-1} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \mathbf{B} \mathbf{X}'_2 \end{bmatrix} . \quad (4.18)$$

c. The following relations are true.

$$\begin{aligned}
\text{i.} \quad \mathbf{A}_{h+}\mathbf{A}_{h1} &= k_+^{-1} \begin{bmatrix} \mathbf{X}_1\mathbf{B}\mathbf{X}'_1 & \mathbf{0} \\ \mathbf{X}_2\mathbf{B}\mathbf{X}'_1 & \mathbf{0} \end{bmatrix} \\
\text{ii.} \quad \mathbf{A}_{h+}\mathbf{A}_{h2} &= k_+^{-1} \begin{bmatrix} \mathbf{0} & \mathbf{X}_1\mathbf{B}\mathbf{X}'_2 \\ \mathbf{0} & \mathbf{X}_2\mathbf{B}\mathbf{X}'_2 \end{bmatrix} \\
\text{iii.} \quad \mathbf{A}_{h+}\mathbf{A}_{h1} + \mathbf{A}_{h+}\mathbf{A}_{h2} &= \mathbf{A}_{h+} \\
\text{iv.} \quad \mathbf{A}_{h+}\mathbf{A}_{h1} + \mathbf{A}_{h1}\mathbf{A}_{h+} &= \mathbf{A}_{h+} + k_+^{-1}k_1\mathbf{A}_{h1} - k_+^{-1}k_2\mathbf{A}_{h2} \\
\text{v.} \quad \mathbf{A}_{h+}\mathbf{A}_{h2} + \mathbf{A}_{h2}\mathbf{A}_{h+} &= \mathbf{A}_{h+} - k_+^{-1}k_1\mathbf{A}_{h1} + k_+^{-1}k_2\mathbf{A}_{h2} \\
\text{vi.} \quad \mathbf{A}_{e+}\mathbf{A}_{e1} &= \mathbf{A}_{e1} \\
\text{vii.} \quad \mathbf{A}_{e+}\mathbf{A}_{e2} &= \mathbf{A}_{e2} \\
\text{viii.} \quad \mathbf{A}_{e+}\mathbf{A}_{ep} &= \mathbf{A}_{ep}
\end{aligned}$$

$$\begin{aligned}
\text{d.} \quad \mathbf{A}_{h+}\mathbf{A}_{e+} &= \mathbf{A}_{h+}\mathbf{A}_{e1} = \mathbf{A}_{h+}\mathbf{A}_{e2} = \mathbf{A}_{h+}\mathbf{A}_{ep} = \mathbf{A}_{h+}\mathbf{A}_{eb} = \mathbf{A}_{h1}\mathbf{A}_{e2} = \mathbf{A}_{h1}\mathbf{A}_{e1} = \\
&\mathbf{A}_{h1}\mathbf{A}_{e2} = \mathbf{A}_{h1}\mathbf{A}_{eb} = \mathbf{A}_{h2}\mathbf{A}_{e1} = \mathbf{A}_{h2}\mathbf{A}_{e2} = \mathbf{A}_{h2}\mathbf{A}_{eb} = \mathbf{A}_{e1}\mathbf{A}_{e2} = \mathbf{A}_{e1}\mathbf{A}_{ep} = \\
&\mathbf{A}_{e1}\mathbf{A}_{eb} = \mathbf{A}_{e2}\mathbf{A}_{ep} = \mathbf{A}_{e2}\mathbf{A}_{eb} = \mathbf{A}_{ep}\mathbf{A}_{eb} = \mathbf{0}_{n_+ \times n_+}
\end{aligned}$$

e. All of the matrices defined in equations 4.4-4.11 are idempotent.

f. For a equal to the number of rows of contrast \mathbf{C} and r the rank of \mathbf{X}_0 , the matrices defined in equations 4.4-4.11 have the following ranks.

Matrix	Rank
\mathbf{A}_{h+}	a
\mathbf{A}_{h1}	a
\mathbf{A}_{h2}	a
\mathbf{A}_{e+}	$n_+ - r$
\mathbf{A}_{e1}	$n_1 - r$
\mathbf{A}_{e2}	$n_2 - r$
\mathbf{A}_{ep}	a
\mathbf{A}_{eb}	$r - a$

A proof for Lemma 4.2 is in Appendix B.

Using the results from Lemma 4.1 and 4.2, I now derive a series of results necessary for the key distributional results of this chapter.

For $s \in \{1, 2, +\}$, \mathbf{A}_{h_s} is idempotent, symmetric, and rank a from Lemma 4.2. Hence \mathbf{V}_{h_s} of dimension $n_+ \times a$ can be defined with

$$\mathbf{A}_{h_s} = \mathbf{V}_{h_s}\mathbf{V}'_{h_s} \tag{4.19}$$

and

$$\mathbf{V}'_{hs} \mathbf{V}_{hs} = \mathbf{I}_a \quad . \quad (4.20)$$

From equation 2.32, it was shown that

$$\mathbf{V}'_{h1} \mathbf{V}_{h+} = \mathbf{V}'_{h+} \mathbf{V}_{h1} = (n_1/n_+)^{1/2} \mathbf{I}_a \quad . \quad (4.21)$$

It can be similarly shown that

$$\mathbf{V}'_{h2} \mathbf{V}_{h+} = \mathbf{V}'_{h+} \mathbf{V}_{h2} = (n_2/n_+)^{1/2} \mathbf{I}_a \quad . \quad (4.22)$$

Also, for

$$\mathbf{V}_{h2} = k_2^{-1/2} \mathbf{X}_{2*} (\mathbf{X}'_0 \mathbf{X}_0)^{-1} \mathbf{C}' \mathbf{F}_0^{-t} \quad , \quad (4.23)$$

the following holds:

$$\begin{aligned} \mathbf{V}'_{h2} \mathbf{V}_{h1} &= (k_2 k_1)^{-1/2} \left[\mathbf{F}_0^{-1} \mathbf{C} (\mathbf{X}'_0 \mathbf{X}_0)^{-1} \mathbf{X}'_{2*} \right] \left[\mathbf{X}_{1*} (\mathbf{X}'_0 \mathbf{X}_0)^{-1} \mathbf{C}' \mathbf{F}_0^{-t} \right] \\ &= (k_2 k_1)^{-1/2} \mathbf{F}_0^{-1} \mathbf{C} (\mathbf{X}'_0 \mathbf{X}_0)^{-1} \begin{bmatrix} \mathbf{0}_{q \times n_1} & \mathbf{X}'_2 \end{bmatrix} \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{0}_{n_2 \times q} \end{bmatrix} (\mathbf{X}'_0 \mathbf{X}_0)^{-1} \mathbf{C}' \mathbf{F}_0^{-t} \\ &= (k_2 k_1)^{-1/2} \mathbf{F}_0^{-1} \mathbf{C} (\mathbf{X}'_0 \mathbf{X}_0)^{-1} \mathbf{0}_{q \times q} (\mathbf{X}'_0 \mathbf{X}_0)^{-1} \mathbf{C}' \mathbf{F}_0^{-t} \\ &= \mathbf{0}_{a \times a} \quad . \end{aligned} \quad (4.24)$$

With

$$\begin{aligned} \begin{bmatrix} \mathbf{V}'_{h1} \\ \mathbf{V}'_{h2} \\ \mathbf{V}'_{h+} \end{bmatrix} \begin{bmatrix} \mathbf{V}_{h1} & \mathbf{V}_{h2} & \mathbf{V}_{h+} \end{bmatrix} &= \begin{bmatrix} \mathbf{I}_a & \mathbf{V}'_{h1} \mathbf{V}_{h2} & \mathbf{V}'_{h1} \mathbf{V}_{h+} \\ \mathbf{V}'_{h2} \mathbf{V}_{h1} & \mathbf{I}_a & \mathbf{V}'_{h2} \mathbf{V}_{h+} \\ \mathbf{V}'_{h+} \mathbf{V}_{h1} & \mathbf{V}'_{h+} \mathbf{V}_{h2} & \mathbf{I}_a \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{I}_a & \mathbf{0} & (n_1/n_+)^{1/2} \mathbf{I}_a \\ \mathbf{0} & \mathbf{I}_a & (n_2/n_+)^{1/2} \mathbf{I}_a \\ (n_1/n_+)^{1/2} \mathbf{I}_a & (n_2/n_+)^{1/2} \mathbf{I}_a & \mathbf{I}_a \end{bmatrix} \end{aligned} \quad (4.25)$$

and, using equation 2.53,

$$\begin{bmatrix} \mathbf{V}'_{h1} \boldsymbol{\mu}_+ \\ \mathbf{V}'_{h2} \boldsymbol{\mu}_+ \\ \mathbf{V}'_{h+} \boldsymbol{\mu}_+ \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_{h1} \\ \boldsymbol{\mu}_{h2} \\ \boldsymbol{\mu}_{h+} \end{bmatrix} = \begin{bmatrix} (n_1/n_+)^{1/2} \boldsymbol{\mu}_{h+} \\ (n_2/n_+)^{1/2} \boldsymbol{\mu}_{h+} \\ \boldsymbol{\mu}_{h+} \end{bmatrix} \quad . \quad (4.26)$$

If

$$\mathbf{y}_h = \begin{bmatrix} \mathbf{V}'_{h1} \\ \mathbf{V}'_{h2} \\ \mathbf{V}'_{h+} \end{bmatrix} \mathbf{y}_+ = \begin{bmatrix} \mathbf{y}_{h1} \\ \mathbf{y}_{h2} \\ \mathbf{y}_{h+} \end{bmatrix}, \quad (4.27)$$

then

$$\mathbf{y}_h \sim (\mathcal{S})\mathcal{N}_{3a} \left(\begin{bmatrix} \boldsymbol{\mu}_{h1} \\ \boldsymbol{\mu}_{h2} \\ \boldsymbol{\mu}_{h+} \end{bmatrix}, \sigma^2 \begin{bmatrix} \mathbf{I}_a & \mathbf{0} & (n_1/n_+)^{1/2} \mathbf{I}_a \\ \mathbf{0} & \mathbf{I}_a & (n_2/n_+)^{1/2} \mathbf{I}_a \\ (n_1/n_+)^{1/2} \mathbf{I}_a & (n_2/n_+)^{1/2} \mathbf{I}_a & \mathbf{I}_a \end{bmatrix} \right). \quad (4.28)$$

Another important relationship is

$$\begin{aligned} & (n_1/n_+)^{1/2} \mathbf{V}'_{h1} + (n_2/n_+)^{1/2} \mathbf{V}'_{h2} = \\ & k_+^{1/2} \mathbf{F}_0^{-1} \mathbf{C}(\mathbf{X}'_0 \mathbf{X}_0)^{-1} \mathbf{X}'_{1*} + k_+^{1/2} \mathbf{F}_0^{-1} \mathbf{C}(\mathbf{X}'_0 \mathbf{X}_0)^{-1} \mathbf{X}'_{2*} = \\ & k_+^{1/2} \mathbf{F}_0^{-1} \mathbf{C}(\mathbf{X}'_0 \mathbf{X}_0)^{-1} [\mathbf{X}'_1 \mathbf{0}_{q \times n_2}] + k_+^{1/2} \mathbf{F}_0^{-1} \mathbf{C}(\mathbf{X}'_0 \mathbf{X}_0)^{-1} [\mathbf{0}_{q \times n_1} \mathbf{X}'_2] = \\ & k_+^{1/2} \mathbf{F}_0^{-1} \mathbf{C}(\mathbf{X}'_0 \mathbf{X}_0)^{-1} [\mathbf{X}'_1 \quad \mathbf{X}'_2] = \mathbf{V}'_{h+} \end{aligned} \quad (4.29)$$

which implies that

$$\mathbf{y}_{h+} = (n_1/n_+)^{1/2} \mathbf{y}_{h1} + (n_2/n_+)^{1/2} \mathbf{y}_{h2}. \quad (4.30)$$

Since \mathbf{y}_{h1} is independent from \mathbf{y}_{h2} , it follows for $a \times 2$ matrix \mathbf{T}_1 defined as

$$\mathbf{T}_1 = [(n_1/n_+)^{1/2} \mathbf{y}_{h1} \quad (n_2/n_+)^{1/2} \mathbf{y}_{h2}] \quad (4.31)$$

that

$$\mathbf{T}_1 \sim \mathcal{N}_{a,2} \{ [(n_1/n_+) \boldsymbol{\mu}_{h+} \quad (n_2/n_+) \boldsymbol{\mu}_{h+}], \mathbf{I}_a, (\sigma^2/n_+) \text{Dg}(n_1, n_2) \}. \quad (4.32)$$

This in turn implies that the 2×2 matrix \mathbf{S}_1 defined as $\mathbf{T}'_1 \mathbf{T}_1$ can be expressed as

$$\begin{aligned} \mathbf{S}_1 &= \begin{bmatrix} S_{11} & S_{12} \\ S_{12} & S_{22} \end{bmatrix} \\ &= \begin{bmatrix} (n_1/n_+)^{1/2} \mathbf{y}'_{h1} \\ (n_2/n_+)^{1/2} \mathbf{y}'_{h2} \end{bmatrix} [(n_1/n_+)^{1/2} \mathbf{y}_{h1} \quad (n_2/n_+)^{1/2} \mathbf{y}_{h2}] \\ &= \begin{bmatrix} (n_1/n_+) \mathbf{y}'_{h1} \mathbf{y}_{h1} & (n_1 n_2 / n_+^2)^{1/2} \mathbf{y}'_{h1} \mathbf{y}_{h2} \\ (n_1 n_2 / n_+^2)^{1/2} \mathbf{y}'_{h1} \mathbf{y}_{h2} & (n_2/n_+) \mathbf{y}'_{h2} \mathbf{y}_{h2} \end{bmatrix} \\ &= \begin{bmatrix} (n_1/n_+) \widehat{\delta}_1 & (n_1 n_2 / n_+^2)^{1/2} \mathbf{y}'_{h1} \mathbf{y}_{h2} \\ (n_1 n_2 / n_+^2)^{1/2} \mathbf{y}'_{h1} \mathbf{y}_{h2} & (n_2/n_+) \widehat{\delta}_2 \end{bmatrix}. \end{aligned} \quad (4.33)$$

Also,

$$\widehat{\delta}_+ = \mathbf{1}'_2 \mathbf{S}_1 \mathbf{1}_2 . \quad (4.34)$$

For 2×2 matrix \mathbf{M}_{S1} defined as

$$\begin{aligned} \mathbf{M}_{S1} &= \begin{bmatrix} (n_1/n_+) \boldsymbol{\mu}'_{h+} \\ (n_2/n_+) \boldsymbol{\mu}'_{h+} \end{bmatrix} \begin{bmatrix} (n_1/n_+) \boldsymbol{\mu}_{h+} & (n_2/n_+) \boldsymbol{\mu}_{h+} \end{bmatrix} \\ &= n_+^{-2} \begin{bmatrix} n_1^2 \boldsymbol{\mu}'_{h+} \boldsymbol{\mu}_{h+} & n_1 n_2 \boldsymbol{\mu}'_{h+} \boldsymbol{\mu}_{h+} \\ n_1 n_2 \boldsymbol{\mu}'_{h+} \boldsymbol{\mu}_{h+} & n_2^2 \boldsymbol{\mu}'_{h+} \boldsymbol{\mu}_{h+} \end{bmatrix} \\ &= n_+^{-2} (\boldsymbol{\mu}'_{h+} \boldsymbol{\mu}_{h+}) \left(\begin{bmatrix} n_1 \\ n_2 \end{bmatrix} \begin{bmatrix} n_1 & n_2 \end{bmatrix} \right) , \end{aligned} \quad (4.35)$$

using the definition of a non-central Wishart variable from Chapter 10 of Muller and Stewart (2006) it follows that

$$\mathbf{S}_1 \sim \mathcal{W}_2[a, (\sigma^2/n_+) \text{Dg}(n_1, n_2), \mathbf{M}_{S1}] . \quad (4.36)$$

If

$$\mathbf{T}_2 = \sigma^{-1} \begin{bmatrix} 1 & (n_2/n_1)^{1/2} \\ 1 & -(n_1/n_2)^{1/2} \end{bmatrix} , \quad (4.37)$$

then

$$\begin{aligned} & \mathbf{T}'_2 (\sigma^2/n_+) \text{Dg}(n_1, n_2) \mathbf{T}_2 = \\ & n_+^{-1} \begin{bmatrix} 1 & 1 \\ (n_2/n_1)^{1/2} & -(n_1/n_2)^{1/2} \end{bmatrix} \begin{bmatrix} n_1 & 0 \\ 0 & n_2 \end{bmatrix} \begin{bmatrix} 1 & (n_2/n_1)^{1/2} \\ 1 & -(n_1/n_2)^{1/2} \end{bmatrix} = \\ & n_+^{-1} \begin{bmatrix} n_1 & n_2 \\ (n_1 n_2)^{1/2} & -(n_1 n_2)^{1/2} \end{bmatrix} \begin{bmatrix} 1 & (n_2/n_1)^{1/2} \\ 1 & -(n_1/n_2)^{1/2} \end{bmatrix} = \\ & n_+^{-1} \begin{bmatrix} n_1 + n_2 & (n_1 n_2)^{1/2} - (n_1 n_2)^{1/2} \\ (n_1 n_2)^{1/2} - (n_1 n_2)^{1/2} & n_2 + n_1 \end{bmatrix} = \\ & n_+^{-1} \begin{bmatrix} n_+ & 0 \\ 0 & n_+ \end{bmatrix} = \mathbf{I}_2 . \end{aligned} \quad (4.38)$$

Also, for $\lambda_+ = \boldsymbol{\mu}'_{h+} \boldsymbol{\mu}_{h+} / \sigma^2$,

$$\begin{aligned}
& \mathbf{T}_2' \mathbf{M}_{S_1} \mathbf{T}_2 = \\
& \frac{(\boldsymbol{\mu}'_{h+} \boldsymbol{\mu}_{h+})}{n_+^2 \sigma^2} \begin{bmatrix} 1 & 1 \\ (n_2/n_1)^{1/2} & -(n_1/n_2)^{1/2} \end{bmatrix} \begin{bmatrix} n_1^2 & n_1 n_2 \\ n_1 n_2 & n_2^2 \end{bmatrix} \begin{bmatrix} 1 & (n_2/n_1)^{1/2} \\ 1 & -(n_1/n_2)^{1/2} \end{bmatrix} = \\
& \frac{(\boldsymbol{\mu}'_{h+} \boldsymbol{\mu}_{h+})}{n_+^2 \sigma^2} \begin{bmatrix} n_1^2 + n_1 n_2 & n_1 n_2 + n_2^2 \\ n_2^{1/2} n_1^{3/2} - n_2^{1/2} n_1^{3/2} & n_2^{3/2} n_1^{1/2} - n_2^{3/2} n_1^{1/2} \end{bmatrix} \begin{bmatrix} 1 & (n_2/n_1)^{1/2} \\ 1 & -(n_1/n_2)^{1/2} \end{bmatrix} = \\
& \frac{(\boldsymbol{\mu}'_{h+} \boldsymbol{\mu}_{h+})}{n_+^2 \sigma^2} \begin{bmatrix} n_1^2 + n_1 n_2 & n_2^2 + n_1 n_2 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & (n_2/n_1)^{1/2} \\ 1 & -(n_1/n_2)^{1/2} \end{bmatrix} = \\
& \frac{(\boldsymbol{\mu}'_{h+} \boldsymbol{\mu}_{h+})}{n_+^2 \sigma^2} \begin{bmatrix} (n_1 + n_2)^2 \left(n_1^{3/2} n_2^{1/2} + n_2^{3/2} n_1^{1/2} \right) - \left(n_2^{3/2} n_1^{1/2} + n_1^{3/2} n_2^{1/2} \right) \\ 0 & 0 \end{bmatrix} = \\
& \frac{(\boldsymbol{\mu}'_{h+} \boldsymbol{\mu}_{h+})}{\sigma^2} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} = \\
& \text{Dg}(\lambda_+, 0) . \tag{4.39}
\end{aligned}$$

The inverse of \mathbf{T}_2 can be expressed as

$$\begin{aligned}
\mathbf{T}_2^{-1} &= -\sigma \left((n_1/n_2)^{1/2} + (n_2/n_1)^{1/2} \right)^{-1} \begin{bmatrix} -(n_1/n_2)^{1/2} & -(n_2/n_1)^{1/2} \\ -1 & 1 \end{bmatrix} \\
&= \sigma n_+^{-1} (n_1 n_2)^{1/2} \begin{bmatrix} (n_1/n_2)^{1/2} & (n_2/n_1)^{1/2} \\ 1 & -1 \end{bmatrix} \\
&= \sigma n_+^{-1} \begin{bmatrix} n_1 & n_2 \\ (n_1 n_2)^{1/2} & -(n_1 n_2)^{1/2} \end{bmatrix} . \tag{4.40}
\end{aligned}$$

For independent random variables

$$X_1^2 \sim \chi^2(a, \lambda_+) , \tag{4.41}$$

$$X_2^2 \sim \chi^2(a - 1) , \tag{4.42}$$

$$Z \sim \mathcal{N}(0, 1) , \tag{4.43}$$

due to the properties of \mathbf{S}_1 and \mathbf{T}_2 described in equations 4.36-4.39, a non-central Wishart theorem (Gupta and Nagar, 2000; Theorem 3.5.8, p.121) allows expressing the distribution of $\mathbf{S}_2 = \mathbf{T}_2' \mathbf{S}_1 \mathbf{T}_2$ in terms of $\{X_1, Z, X_2\}$. In particular,

$$\begin{aligned}
\mathbf{S}_2 &= \mathbf{T}'_2 \mathbf{S}_1 \mathbf{T}_2 \\
&= \begin{bmatrix} X_1 & 0 \\ Z & X_2 \end{bmatrix} \begin{bmatrix} X_1 & Z \\ 0 & X_2 \end{bmatrix} \\
&= \begin{bmatrix} X_1^2 & X_1 Z \\ X_1 Z & Z^2 + X_2^2 \end{bmatrix}.
\end{aligned} \tag{4.44}$$

This in turn implies that \mathbf{S}_1 can be expressed as

$$\begin{aligned}
\mathbf{S}_1 &= \mathbf{T}_2^{-t} \mathbf{S}_2 \mathbf{T}^{-1} \\
&= \sigma^2 n_+^{-2} \begin{bmatrix} n_1 & (n_1 n_2)^{1/2} \\ n_2 & -(n_1 n_2)^{1/2} \end{bmatrix} \begin{bmatrix} X_1^2 & X_1 Z \\ X_1 Z & Z^2 + X_2^2 \end{bmatrix} \begin{bmatrix} n_1 & n_2 \\ (n_1 n_2)^{1/2} & -(n_1 n_2)^{1/2} \end{bmatrix}.
\end{aligned} \tag{4.45}$$

So, for $\mathbf{S}_1 = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}$, the components of S_1 are

$$\begin{aligned}
S_{11} &= \sigma^2 n_+^{-2} \left[n_1^2 X_1^2 + 2n_1 (n_1 n_2)^{1/2} X_1 Z + n_1 n_2 (Z^2 + X_2^2) \right] \\
&= \sigma^2 n_+^{-2} n_1 \left[\left(n_1^{1/2} X_1 + n_2^{1/2} Z \right)^2 + n_2 X_2^2 \right]
\end{aligned} \tag{4.46}$$

$$\begin{aligned}
S_{12} &= \sigma^2 n_+^{-2} \left[n_1 n_2 X_1^2 + (n_2 - n_1) (n_1 n_2)^{1/2} X_1 Z - n_1 n_2 (Z^2 + X_2^2) \right] \\
&= \sigma^2 n_+^{-2} \left[n_1 n_2 (X_1^2 - Z^2 - X_2^2) + (n_2 - n_1) (n_1 n_2)^{1/2} X_1 Z \right]
\end{aligned} \tag{4.47}$$

$$\begin{aligned}
S_{21} &= \sigma^2 n_+^{-2} \left[n_1 n_2 X_1^2 + (n_2 - n_1) (n_1 n_2)^{1/2} X_1 Z - n_1 n_2 (Z^2 + X_2^2) \right] \\
&= \sigma^2 n_+^{-2} \left[n_1 n_2 (X_1^2 - Z^2 - X_2^2) + (n_2 - n_1) (n_1 n_2)^{1/2} X_1 Z \right]
\end{aligned} \tag{4.48}$$

$$\begin{aligned}
S_{22} &= \sigma^2 n_+^{-2} \left[n_2^2 X_2^2 - 2n_2 (n_1 n_2)^{1/2} X_1 Z + n_1 n_2 (Z^2 + X_2^2) \right] \\
&= \sigma^2 n_+^{-2} n_2 \left[\left(n_2^{1/2} X_1 - n_1^{1/2} Z \right)^2 + n_1 X_2^2 \right].
\end{aligned} \tag{4.49}$$

Now, using the results above, I can get desired expressions for the conditional sums of squares hypothesis for multiple degree of freedom hypotheses.

Lemma 4.3 For $a > 1$ and $X_1^2 \sim \chi^2(a, \lambda_+)$ defined as in equation 4.41, the conditional sum of squares hypothesis for the final test statistic can be written as

$$\widehat{\delta}_+ = \sigma^2 X_1^2. \quad (4.50)$$

A proof for Lemma 4.3 is in Appendix B.

Corollary 4.1 For $a > 1$ and $X_1^2 \sim \chi^2(a, \lambda_+)$ defined as in equation 4.41, the conditional sum of squares hypothesis for the final test statistic, scaled by σ^2 , can be written as

$$G_+ = \widehat{\delta}_+ / \sigma^2 = X_1^2. \quad (4.51)$$

Proof. Follows directly from equation 4.50.

Lemma 4.4 For $a > 1$ and X_1^2 , X_2^2 , and Z defined as in equations 4.41-4.43, the conditional sum of squares hypothesis for the first test statistic can be written as

$$\widehat{\delta}_1 = \sigma^2 n_+^{-1} \left[\left(n_1^{1/2} X_1 + n_2^{1/2} Z \right)^2 + n_2 X_2^2 \right]. \quad (4.52)$$

A proof for Lemma 4.4 is in Appendix B.

Corollary 4.2 For $a > 1$ and X_1^2 , X_2^2 , and Z defined as in equations 4.41-4.43, the conditional sum of squares hypothesis for the first test statistic, scaled by σ^2 , can be written as

$$G_1 = \widehat{\delta}_1 / \sigma^2 = n_+^{-1} \left[\left(n_1^{1/2} X_1 + n_2^{1/2} Z \right)^2 + n_2 X_2^2 \right]. \quad (4.53)$$

Proof. Follows directly from equation 4.52.

Lemma 4.5 For $a > 1$ and X_1^2 , X_2^2 , and Z defined as in equations 4.41-4.43, the conditional sum of squares hypothesis from the second stage can be written as

$$\widehat{\delta}_2 = \sigma^2 n_+^{-1} \left[\left(n_2^{1/2} X_1 - n_1^{1/2} Z \right)^2 + n_1 X_2^2 \right]. \quad (4.54)$$

A proof for Lemma 4.5 is in Appendix B.

Corollary 4.3 For $a > 1$ and X_1^2 , X_2^2 , and Z defined as in equations 4.41-4.43, the conditional sum of squares hypothesis from the second stage, scaled by σ^2 , can be written as

$$G_2 = \widehat{\delta}_2/\sigma^2 = n_+^{-1} \left[\left(n_2^{1/2} X_1 - n_1^{1/2} Z \right)^2 + n_1 X_2^2 \right]. \quad (4.55)$$

Proof. Follows directly from equation 4.54.

Corollary 4.4 For $a > 1$, X_1^2 , X_2^2 , and Z defined as in equations 4.41-4.43, and G_1 and G_2 defined in equations 4.53 and 4.55, the following holds true

$$G_1 + G_2 = \left(\widehat{\delta}_1 + \widehat{\delta}_2 \right) / \sigma^2 = X_1^2 + X_2^2 + Z^2. \quad (4.56)$$

Proof. Follows directly from summing the results in equations 4.53 and 4.55.

4.3 THE IPIA PROCEDURE AND PROPERTIES

Table 4.1: *General procedure*

Step 1a : Specify α_t , P_t , \mathbf{X}_0 , hypotheses, $\boldsymbol{\theta}_1$, and σ_0^2
1b : Solve for first stage sample size (n_1)
Step 2 : Collect first n_1 observations
Step 3 : Solve for $N_+ = n_+$, critical values $f_l(n_+)$, $f_u(n_+)$, and $f_+(n_+)$, and F_1
Step 4 : Decide:
If $F_1 < f_l$ then STOP, ACCEPT H_0
If $F_1 \geq f_u$ then STOP, REJECT H_0
If $f_l \leq F_1 < f_u$ then take $n_2 = n_+ - n_1$ additional observations
Step 5 : Solve for F_+
Step 6 : Decide:
If $F_+ < f_+$ then ACCEPT H_0
If $F_+ \geq f_+$ then REJECT H_0

Table 4.1 outlines the general procedure for the IPIA model. The order of the steps matters in specifying the distributions.

The value of the internal pilot sample size, n_1 , must be chosen at the design stage of the study. The choice is important since lower values give more uncertain estimates of σ^2 while higher values reduce possible savings in sample size. Most authors discussing internal pilot designs take a designated fraction of the sample size from fixed sample equations such as $n_1 = \pi \cdot n_0$ for $0 < \pi \leq 1$ and n_0 determined from σ_0^2 . A typical choice for π seems to be

0.5; that is, the size of the first sample is half of the fixed sample study sample size based on σ_0^2 . In Chapter 3, I showed that this choice is a complex decision and should be a design factor during study planning with the goals of an individual study in mind. In order to clearly portray the usage and properties of the IPIA technique compared to other design types for multiple degree of freedom hypotheses, I will not enumerate examples for various values of π here. This will be undertaken as part of additional research into design strategies for multiple degree of freedom hypothesis tests (Chapter 5). For the examples considered in this chapter, the value as close to possible to $\pi = 0.5$ with conforming sample size will be used.

Calculation of the three critical values for the study, $f_l(n_+)$, $f_u(n_+)$, and $f_+(n_+)$, must be done following rules pre-specified in the study protocol. The critical values may depend on n_+ , the realized value of N_+ ; however, when it is clear, they will be referred to as f_l , f_u , and f_+ . Ideally, they should be chosen in a way that controls the type I error rate while having good power and expected sample size properties. The theory developed here optionally allows for stopping under the null at the interim analysis if $F_1 < f_l$, where f_l is the first stage lower critical value. This can cause a great reduction in expected sample size when the effect size is near the null value by allowing the study to stop for a "lost cause". If early stopping under the null is not allowed, then $f_l = 0$ for all $n_+ \neq n_1$. In all cases $f_l = f_u$ when $n_+ = n_1$, which guarantees stopping for acceptance or rejection of the null. Detailed exploration and comparison of sample size selection methods will be saved for future research (Chapter 5).

The sample size re-estimation rule will determine the distribution of N_+ . It is an important consideration in the design affecting type I error rate, power, and expected sample size. Like internal pilot designs, the sample size for IPIA designs is determined by using the updated variance estimate at the interim stage to recalculate the estimated sample size need to achieve target power in the final test. The procedure takes advantage of the monotone relationship between continuous $\hat{\sigma}_1^2$ and discrete N_+ .

Computing the distribution of sample size, the cumulative distribution function for possible values of N_+ , that is $\Pr\{N_+ \leq n\}$, requires determining cut-off points based on the first stage variance estimate. For a particular value n_+ of random N_+ , the first step is to solve for scaled noncentrality $\lambda(n_+)$ that satisfies what about notation $\lambda_{\min}(n_+)$

$$P_t = 1 - F_{\chi^2}[f_{\text{crit}}; a, \lambda(n_+)] \quad (4.57)$$

or

$$P_t = 1 - F_F[f_{\text{crit}}; a, \nu_+, \lambda(n_+)] , \quad (4.58)$$

with $f_{\text{crit}} = F_{\chi^2}^{-1}(1 - \alpha_t; a)$ or $F_F^{-1}(1 - \alpha_t; a, \nu_+)$ depending on whether or not large sample distributional assumptions are used. The scaled noncentrality $\lambda(n_+)$ here represents the minimum value that would lead to a final sample size of n_+ . Since the effect of interest, θ_1 , is used at the planning stage, for $\delta(n_+) = \theta_1' \mathbf{M}_+^{-1} \theta_1$ the following hold:

$$\lambda(n_+) = \delta(n_+)/\sigma^2(n_+) \quad (4.59)$$

or

$$\sigma^2(n_+) = \delta(n_+)/\lambda(n_+) . \quad (4.60)$$

Here, $\sigma^2(n_+)$ designates the largest value of $\hat{\sigma}_1^2$ that would produce $N_+ = n_+$. Therefore, since $\nu_1 \hat{\sigma}_1^2 \sim \sigma^2 W$ for $W \sim \chi^2(\nu_1)$,

$$\begin{aligned} \Pr\{N_+ \leq n_+\} &= \Pr\{\hat{\sigma}_1^2 \leq \sigma^2(n_+)\} \\ &= \Pr\{W \leq \nu_1 \sigma^2(n_+)/\sigma^2\} \\ &= \Pr\{W \leq \nu_1 \delta(n_+)/[\sigma^2 \lambda(n_+)]\} . \end{aligned} \quad (4.61)$$

The discreteness of sample size implies

$$\Pr\{N_+ = n_+\} = \Pr\{N_+ \leq n_+\} - \Pr\{N_+ \leq n_+ - m\} . \quad (4.62)$$

When restrictions are given for minimum or maximum sample size, the tail probabilities are collapsed into the smallest or largest allowable values, respectively.

A key result of this process is the determination of cut-off points that determine a range into which continuous $\hat{\sigma}_1^2$ must have fallen in order for a given final sample size to occur.

Define $q_1(n_+)$ and $q_2(n_+)$ to be the values such that

$$N_+ = n_+ \Leftrightarrow q_1(n_+) < \nu_1 \hat{\sigma}_1^2 / \sigma^2 \leq q_2(n_+), \quad (4.63)$$

which in turn implies that

$$\Pr\{N_+ = n_+\} = F_{\chi^2}[q_2(n_+); \nu_1] - F_{\chi^2}[q_1(n_+); \nu_1]. \quad (4.64)$$

The cut off points determine the probabilities for discrete values of N_+ and hence describe the variable's distribution. When it is unambiguous, q_1 and q_2 are used for $q_1(n_+)$ and $q_2(n_+)$.

4.4 KEY ANALYTIC RESULTS FOR PROCEDURE

The results in this paper are developed to test $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ for a contrast matrix \boldsymbol{C} with more than one row, $a > 1$. Without loss of generality I assume $\boldsymbol{\theta}_0 = \mathbf{0}$ (by Lemma A.1).

In order to compute overall power and type I error rate for a study design, a joint distribution of the two stage test statistics is necessary. Since critical values and denominator degrees of freedom depend on sample sizes, the joint CDF is derived conditional on an N_+ value, or $F_{F_1, F_+ | N_+}(f_1, f_+)$ where F_1 and F_+ are the test statistics for the first and second (final) stage, respectively. Using the law of total probability and summing over possible values for sample size gives unconditional power.

Unconditional power requires computing each conditional result. The conditional joint CDF of the test statistics, $F_{F_1, F_+ | N_+}(f_1, f_+)$, is derived in a computable form by decomposing the parts into independent elements. Lemma 4.6 gives a key form involving three independent variables defined in Section 4.2.3.

Lemma 4.6 Intermediate functions of interest include $u_1 = c_+ g_+$, $u_2 = c_+(g_1/n_2 + g_+)$, $u_3(p_1) = c_+^{-1} p_1 - g_+$, $u_4 = g_1/n_2$, $b(p_1, p_2) = \sqrt{g_1/n_2 - p_2} - \sqrt{p_1 n_1/n_2}$, $f_{p_1} = f_{X_1^2}(p_1)$, $f_{p_2} = f_{X_2^2}(p_2)$, $d(p_1, p_2) = \sqrt{g_1/n_2 - p_2} + \sqrt{p_1 n_1/n_2}$, $h(p_1, p_2) = \sqrt{c_+^{-1} p_1 - p_2 - g_+}$. For strictly positive $\{g_1, g_+, c_+, n_1, n_2\}$, integer $a > 1$, $Z \sim \mathcal{N}(0, 1)$, $X_1^2 \sim \chi^2(a, \lambda_+)$, $X_2^2 \sim \chi^2(a - 1)$ with $\{Z, X_1^2, X_2^2\}$ mutually independent,

the following holds.

$$\begin{aligned}
\Pr \left\{ \left[(\sqrt{n_1}X_1 + \sqrt{n_2}Z)^2 + n_2X_2^2 \leq g_1 \right] \cap [c_+^{-1}X_1^2 - X_2^2 - Z^2 \leq g_+] \right\} = \\
\int_0^{u_1} \int_0^{u_4} f_{p_1} f_{p_2} \{ \Phi[b(p_1, p_2)] - \Phi[-d(p_1, p_2)] \} dp_2 dp_1 + \\
\int_{u_1}^{u_2} \int_0^{u_3(p_1)} f_{p_1} f_{p_2} \left[\max \left(\Phi \{ \min[-h(p_1, p_2), b(p_1, p_2)] \} - \Phi[-d(p_1, p_2)], 0 \right) + \right. \\
\left. \max \{ \Phi[b(p_1, p_2)] - \Phi[h(p_1, p_2)], 0 \} \right] dp_2 dp_1 + \\
\int_{u_1}^{u_2} \int_{u_3(p_1)}^{u_4} f_{p_1} f_{p_2} \{ \Phi[b(p_1, p_2)] - \Phi[-d(p_1, p_2)] \} dp_2 dp_1 + \\
\int_{u_2}^{\infty} \int_0^{u_4} f_{p_1} f_{p_2} \left[\max \left(\Phi \{ \min[-h(p_1, p_2), b(p_1, p_2)] \} - \Phi[-d(p_1, p_2)], 0 \right) + \right. \\
\left. \max \{ \Phi[b(p_1, p_2)] - \Phi[h(p_1, p_2)], 0 \} \right] dp_2 dp_1 . \tag{4.65}
\end{aligned}$$

A proof for Lemma 4.6 is in Appendix B.

The result from Lemma 4.6 can be used in the following result in order to calculate the conditional joint distribution of the test statistics. From Coffey and Muller (1999), conditional on N_+ , $E_1 = \nu_1 \hat{\sigma}_1^2 / \sigma^2 \sim \chi_T^2(\nu_1; q_1, q_2)$. *The following theorem provides an explicit form for the desired conditional CDF.*

Theorem 4.1 If $Z \sim \mathcal{N}(0, 1)$, $X_1^2 \sim \chi^2(a, \lambda_+)$, and $X_2^2 \sim \chi^2(a - 1)$ with $\{Z, X_1^2, X_2^2\}$ mutually independent. Additionally, define $E_+ = \nu_+ \hat{\sigma}_+^2 / \sigma^2 = \mathbf{y}'_+ \mathbf{A}_{e1} \mathbf{y}_+ / \sigma^2$, $E_1 = \nu_1 \hat{\sigma}_1^2 / \sigma^2 = \mathbf{y}'_+ \mathbf{A}_{e1} \mathbf{y}_+ / \sigma^2$, $E_2 = \mathbf{y}'_+ (\mathbf{A}_{e2} + \mathbf{A}_{eb}) \mathbf{y}_+ / \sigma^2$, $E_3 = \mathbf{y}'_+ \mathbf{A}_{ep} \mathbf{y}_+ / \sigma^2$, $G_s = \hat{\delta}_s / \sigma^2 = \mathbf{y}'_+ \mathbf{A}_{hs} \mathbf{y}_+ / \sigma^2$, $c_s = a f_s / \nu_s$ for $s \in \{1, 2, +\}$, $f_{E_1}(t_1) = f_{\chi^2}(t_1; \nu_1) / [F_{\chi^2}(q_2; \nu_1) - F_{\chi^2}(q_1; \nu_1)]$, and $f_{E_2}(t_2) = f_{\chi^2}(t_2; n_2 - a)$, then for $a > 1$, the conditional joint distribution of the two stage test statistics can be written as

$$\begin{aligned}
F_{F_1, F_+ | N_+}(f_1, f_+) = \\
\int_{q_1}^{q_2} \int_0^{\infty} f_{E_1}(t_1) f_{E_2}(t_2) \Pr_{N_+} \left\{ \left[(\sqrt{n_1}X_1 + \sqrt{n_2}Z)^2 + n_2X_2^2 \leq n_+ c_1 t_1 \right] \right. \\
\left. \cap [c_+^{-1}X_1^2 - X_2^2 - Z^2 \leq t_1 + t_2] \right\} dt_2 dt_1 \tag{4.66}
\end{aligned}$$

A proof for Theorem 4.1 is in Appendix B.

A result needed if early futility stopping is allowed follows directly from Theorem 4.1:

$$\Pr\{f_l \leq F_1 \leq f_u, F_+ \leq f_+ | N_+ = n_+\} = F_{F_1, F_+ | N_+}(f_u, f_+) - F_{F_1, F_+ | N_+}(f_l, f_+) . \quad (4.67)$$

The results from Theorems 4.1 and 4.2 can be used to explicitly solve for the conditional joint distribution of the test statistics for multiple degree of freedom hypotheses in the IPIA setting. Taken together with results already derived in Chapter 2, explicit calculations can be made for power, type I error, and expected sample size. The following results from Chapter 2 apply to the $a > 1$ setting by using the form for $F_{F_1, F_+ | N_+}(f_1, f_+)$ derived in Theorem 4.1.

A distribution important to power and expected size calculations is the CDF of the first test statistic, conditional on N_+ , i.e., $F_{F_1 | N_+}(f_1)$.

Theorem 4.2 For $\lambda_1 = \boldsymbol{\mu}'_{h1} \boldsymbol{\mu}_{h1} / \sigma^2$, the conditional CDF of the first test statistic can be written

$$F_{F_1 | N_+}(f_1) = \int_{q_1}^{q_2} \frac{F_{\chi^2}(c_1 t_1; a, \lambda_1) f_{\chi^2}(t_1; \nu_1)}{F_{\chi^2}(q_2; \nu_1) - F_{\chi^2}(q_1; \nu_1)} dt_1 . \quad (4.68)$$

A proof of Theorem 4.2 is in Appendix A.

The following corollary adapts Theorem 4.2 to solve for the probability of the test continuing to the second stage conditional on $N_+ = n_+$ when futility stopping is possible.

Corollary 4.5

$$\Pr\left\{f_l \leq F_1 < f_u \mid N_+ = n_+\right\} = \int_{q_1}^{q_2} \frac{[F_{\chi^2}(c_u t_1; a, \lambda_1) - F_{\chi^2}(c_l t_1; a, \lambda_1)] f_{\chi^2}(t_1; \nu_1)}{F_{\chi^2}(q_2; \nu_1) - F_{\chi^2}(q_1; \nu_1)} dt_1 . \quad (4.69)$$

The proof parallels the proof of Theorem 4.2, in Appendix A.

The above results can be used to calculate exact expressions for the power, type I error rate (power under the null hypothesis), and expected sample size. The values change with

design parameters and are valuable knowledge in study planning. *The following theorem gives the formula for unconditional power.*

Theorem 4.3 An expression for unconditional power, P_w , can be written

$$P_w = 1 - \sum_{\{N_+=n_+\}} [F_{\chi^2}(q_2; \nu_1) - F_{\chi^2}(q_1; \nu_1)] \left\{ F_{F_1|N_+}(f_l(n_+)) + \Pr \left[(f_l(n_+) \leq F_1 < f_u(n_+), F_+ < f_+(n_+)) \mid N_+ = n_+ \right] \right\}. \quad (4.70)$$

A proof of Theorem 4.3 is in Appendix A.

The results in this section can also be applied to calculate an expected sample size formula for a study design in the following form.

Theorem 4.4 If N_w equals the total sample size taken in study, that is,

$$N_w = \begin{cases} n_1 & \text{if study stopped after first stage} \\ N_+ & \text{otherwise,} \end{cases}$$

then

$$E(N_w) = n_1 + \sum_{\{N_+=n_+\}} n_2 [F_{\chi^2}(q_2; \nu_1) - F_{\chi^2}(q_1; \nu_1)] \Pr \left[f_l \leq F_1 < f_u \mid N_+ = n_+ \right] \quad (4.71)$$

A proof of Theorem 4.4 is in Appendix A.

4.5 AN EXAMPLE

4.5.1 Motivation for the Example

To illustrate usage of the exact theory in the application of internal pilot studies to more complex designs, I consider Example 4.1, a three-group one-way analysis of variance (ANOVA) described by Coffey and Muller (1999; Example C) for use in internal pilot designs. There are two main purposes of the example in this chapter. First, to compare the numeric results using exact theory to simulations in order to justify the computational algorithms utilized in calculations and as an additional check on the accuracy of the theory. Second, to compare the properties of the internal pilot with interim analysis (IPIA) procedure

with special cases, including a fixed sample, an internal pilot (IP), and a two-stage group sequential (GS) procedure in a multiple degree of hypothesis setting. To help simplify the comparisons, I use a naive but common approach to critical value selection and study design. Properties to be examined include type I error rate, power, and expected sample sizes under various scenarios.

The fixed sample, IP and two-stage GS designs considered for Example 4.1 are all special cases within the general IPIA framework. An IP design does not allow early stopping at the interim power analysis, except when $N_2 = 0$. An IP design may be described as special case of IPIA with $f_l = 0, f_u = \infty$. In the IPIA design used here, it is assumed that sample size can be reduced from the pre-planned level ($n_{+,min} = n_1$). The two-stage GS design, on the other hand, allows for early stopping at the interim analysis, but does not allow for a change to the preplanned maximum sample size, i.e., $\Pr\{N_+ = n_0\} = 1$. The fixed sample approach can be seen as a special case combining the restrictions of the IP and GS designs ($f_l = 0, f_u = \infty$, and $\Pr\{N_+ = n_0\} = 1$). The IPIA design combines the features of the IP and two-stage GS designs by allowing for stopping at the interim stage as well as allowing for a change in maximum sample size used when a study is to be continued.

Table 4.2: *Two-stage designs*
Early Stopping

		Yes	No
SSR	Yes	IPIA	Int Pilot
	No	Grp Seq	Fixed Sample

In addition to stopping at the interim stage for efficacy, both the GS and the IPIA procedures can allow early stopping at the interim stage for futility. Hence both scenarios will be considered here. No futility stopping implies that the lower first stage critical value, f_l , is 0. For futility stopping in this chapter, I will use a simple p-value cut point of 0.85. Therefore if the p-value for the first stage hypothesis test is greater than 0.85, the study will stop and conclude that the alternative hypothesis is not supported. In reality, this is not an

ideal approach since the first stage may contain only a small fraction of the needed information of the study (especially for high true variance values) and so may not be very informative in some cases. It is used here for simplicity in order to portray the characteristics of the procedures. Chapter 5 has some additional discussion about the use of futility stopping within the IPIA framework.

Critical values used in the example will be determined as follows. For the fixed sample, group sequential, and IPIA designs critical values will be based on the chi-square distribution with a degrees of freedom. For example, if a particular stage's nominal alpha level is determined to be $\alpha = 0.04$, then the efficacy critical value for that stage would be $F_{\chi^2}^{-1}(1 - 0.04; a)/a$. The large sample method is used to show the consequences of not accounting for the use of variance estimates in the test statistics for the small and moderate studies examined. The fixed sample results will also be determined using the F distribution since it will exactly achieve the target type I error rate. For the internal pilot design, critical values will only be solved with the unadjusted F distribution since it is the method described by literature. For these methods, a particular stage nominal alpha level of $\alpha = 0.04$ would cause the efficacy critical value for that stage (say stage s) to be $F_F^{-1}(1 - 0.04; a, \nu_s)$. For the group sequential and IPIA designs (early stopping designs), I will use O'Brien-Fleming stopping rules to solve for the nominal type I error rates used in critical value calculation. These bounds are designed to allow for conservative early stopping while adjusting the final critical value for type I error rate inflation due to multiple testing.

Due to misalignment of test statistic and critical value distributions and a biased variance value used in sample size re-estimation designs, the selection of final critical values will likely cause type I error rate inflation for most designs considered. They are used here in order to best compare the procedures and also to assess the magnitudes of such inflations. Strategies that better control the type I error rate and maintain power while minimizing

expected sample size for these designs are saved for future research and will be briefly discussed in Chapter 5.

In total seven design procedures will be considered: fixed sample (χ^2 and F), IP, two-stage GS with and without futility stopping, and IPIA with and without futility stopping. Type I error rate, power, and expected sample size will be calculated for each of these procedures over a range of true variances. I will use the sampling fraction $\pi = 0.5$ and $n_{+, \max} = \infty$.

4.5.2 Computational Methods

All programs for the example were written in SAS/IML (SAS Institute, 2004). Most of the computation for the examples utilizes the exact theory developed in this chapter. The exceptions are the fixed sample and internal pilot designs. The fixed results could be directly calculated using standard distribution functions. The internal pilot calculations were easily obtained utilizing exact internal pilot theory from the freely available GLUMIP 2.0 (Kairalla et al. 2007) software package. All other results came from use of the exact theory, including the two-stage group sequential designs, which are a special case.

Stopping bound computation utilized the SEQSCALE function and the numeric integrations utilized the QUAD function, both within SAS/IML. To avoid numerical instability of the calculated integrals, computation was performed using quantile transformations (Glueck and Muller, 2001) of the distributions derived in Section 4.4. For illustration, transforming variable of integration t to give $p = F_{\chi^2}(t; \nu)$ implies $t = F_{\chi^2}^{-1}(p; \nu)$ and $dp = f_{\chi^2}(t; \nu)dt$. The approach always gives finite bounds and often radically improves computational accuracy and speed.

Simulations were conducted for a limited set of cases in order to check the accuracy of the programming and numerical algorithms, provide an additional check on the analytical derivations, and to compare the speed of calculation using the two methods. Using a subset of a half dozen cases from over a range of conditions, simulation was conducted with

1,000,000 replications per case. All programs were run using an Intel Xeon 3.2 GHz processor. For each of the cases considered, the analytically calculated values were within two standard deviations of the simulated values.

The comparison programs were each run in groups of three cases corresponding to variance values of $\gamma \in \{0.5, 1, 2\}$ with $\gamma = \sigma^2/\sigma_0^2$. Runs were made under the null hypothesis ($\boldsymbol{\theta} = \mathbf{0}$) and assuming the effect of interest ($\boldsymbol{\theta} = \boldsymbol{\theta}_1$) for the IPIA design without futility for Example 4.1. Timing results are detailed in Table 4.3 below.

Table 4.3: *Simulation and calculation times (hours) for Ex. 4.1*

	Simulation	Calculation
$\boldsymbol{\theta} = \mathbf{0}$	2.9	16.2
$\boldsymbol{\theta} = \boldsymbol{\theta}_1$	2.8	17.3

For the examples considered, the analytic calculations using the exact theory were slower than the simulations (about 5x). This is not surprising considering the program calculates four dimensional integrals as nested univariate integrations using the QUAD functions in SAS/IML and no effort has been made thus far to improve the computational efficiency of the program. Given the speed improvements seen in the very similar $a = 1$ case, even these complex results should be more efficient than simulations.

4.5.3 Example 4.1 Results

Example 4.1 is a three-group one-way analysis of variance (ANOVA) example previously described by Coffey and Muller (1999; Example C) in an internal pilot framework. For the two degree of freedom test of differences among groups with $\alpha_t = 0.05$, $P_t = 0.90$, $\boldsymbol{\theta}_1 = [0.5 \quad 1.0]'$, and $\sigma_0^2 = 1$, a fixed sample size power calculation suggests 27 observations per group ($n_0 = 81$). For the early stopping procedures, I consider a design with 13 observations per group ($n_1 = 39$) in the interim sample and $n_{+, \max} = \infty$. The design parameters for Example 4.1 are summarized in Table 4.4.

Table 4.4: *Design parameters for Example 4.1*

α_t	P_t	θ_1	σ_0^2	n_0	n_1	$n_{+,max}$
0.05	0.9	[0.5 1.0]'	1	81	39	∞

I analytically calculated values for type I error rate, power, and expected sample size under the design conditions described in Table 4.4.

Table 4.5: *Type I error rates $\times 100$ for Example 4.1*

γ	Fixed Sample		IP	Group Sequential		IPIA	
	χ^2	F		w/o Futility	w/ Futility	w/o Futility	w/ Futility
0.50	5.6	5.0	5.3	5.9	5.8	6.4	6.4
0.75	5.6	5.0	5.6	5.9	5.8	6.8	6.8
1.00	5.6	5.0	5.5	5.9	5.8	6.6	6.6
1.50	5.6	5.0	5.3	5.9	5.8	6.1	5.9
2.00	5.6	5.0	5.2	5.9	5.8	5.7	5.4

Table 4.5 displays the values for type I error rate for each of the seven designs described in Section 4.5.1: Fixed sample (χ^2 and F), IP, two-stage GS with and without futility stopping, and IPIA with and without futility stopping.

For the fixed sample design, type I error rate is somewhat inflated by a constant amount across γ when the large sample chi-square distribution is used for critical value determination and is controlled at the target level when the F distribution is used. In the IP design, some type I error rate inflation occurs due to downward bias in variance estimate used. The magnitude of inflation is shown to depend on true variance value with a peak at around $\gamma = 0.75$. Due to the use of chi-square based critical values, the GS designs also have moderate type I error rate inflation. The inflation for the GS designs is constant across γ since no sample size re-estimation occurs and noncentrality is zero under the null. The IPIA designs, which combine early stopping ability with sample size re-estimation, have type I error rate inflation caused by both variance estimate bias and the use of large sample critical values. For both GS and IPIA, allowing for early stopping for futility causes a small reduction in the type I error rate. For this example, the magnitude of type I error rate is

moderate and comparable for the Fixed (χ^2), IP, and GS methods. The IPIA method has an increased level of inflation.

Table 4.6: $Power \times 100$ for Example 4.1

γ	Fixed Sample			Group Sequential		IPIA	
	χ^2	F	IP	w/o Futility	w/ Futility	w/o Futility	w/ Futility
0.50	99.8	99.7	93.3	99.8	99.7	93.3	93.3
0.75	97.3	96.9	91.2	97.2	97.1	90.9	90.8
1.00	91.6	90.8	90.4	91.5	91.2	90.2	90.0
1.50	76.9	75.4	89.6	76.8	76.4	89.5	88.3
2.00	63.9	62.1	89.1	63.9	63.5	89.1	86.8

Table 4.5 displays the values of unconditional power for the seven designs. Power for both fixed sample designs is sensitive to the true variance value. The fixed sample study considered can be significantly over or under powered depending on the true variance regardless of the critical value determination method employed. Power for the considered GS designs is also highly dependent on the true variance value, with power levels very similar to those for the fixed sample design. In the IP design, power is greatly stabilized due to the variance estimate based sample size re-estimation at first stage. The IPIA designs also have very stable power similar to the IP design due to the sample size re-estimation analysis at the first stage. The GS and IPIA designs with first stage futility stopping have power at slightly lower levels than their counterparts without futility stopping. For power in this example, the IP and IPIA designs greatly achieve the target rate while the two-stage GS and fixed sample designs are shown to be vulnerable to misspecification of the variance, a nuisance parameter, at the planning stage.

Table 4.7: $E(N_w)$ for Example 4.1: fixed, IP, and GS

γ	Fixed Sample	IP	GS (no futility)			GS (futility)		
			$\theta = 0$	$\theta = \theta_1$	$\theta = 2\theta_1$	$\theta = 0$	$\theta = \theta_1$	$\theta = 2\theta_1$
0.50	81	44.5	80.6	52.7	39.0	74.4	52.7	39.0
0.75	81	61.6	80.6	62.5	39.2	74.4	62.4	39.2
1.00	81	80.5	80.6	68.1	40.5	74.4	67.8	40.5
1.50	81	118.4	80.6	73.5	46.1	74.4	72.6	46.1
2.00	81	156.4	80.6	75.9	52.7	74.4	74.5	52.7

Table 4.7 displays the values for expected sample size for the fixed sample, IP, and GS with and without futility stopping designs. For the GS designs, expected sample sizes are calculated assuming the null hypothesis ($\theta = 0$), the alternative of interest ($\theta = \theta_1$), and assuming a true effect size twice the effect of interest ($\theta = 2\theta_1$).

Under controlled conditions, the sample size for the fixed sample design is always the preplanned sample size, 81. As would be expected, the expected sample size for the IP design is dependent on the true variance due to sample estimate based sample size re-estimation at the first stage. It achieves an expected savings in sample size for variance values lower than the value assumed at the planning stage (44.5 at $\gamma = 0.5$) and rises above that of the fixed sample design as it accounts for larger true variance values by increasing the estimated sample size need at the internal pilot stage (156.4 at $\gamma = 2.0$).

Under the null hypothesis, the expected sample size for the GS designs are constant over γ values. This happens because no variance-value based sample size re-estimation takes place at the first stage and true noncentrality is zero. The small departure from the fixed design sample size in the GS design without futility stopping is due to the small chance of falsely stopping for efficacy at the first stage. The GS designs allowing futility stopping at the first stage causes an across the board drop in expected sample size (to 74.4 from 80.6) under the null due the probability of correctly stopping early for futility.

Under the alternative of interest ($\theta = \theta_1$), the expected sample sizes for the GS designs are noticeable lower than the fixed design sample sizes due to possible early stopping for efficacy at first stage. This demonstrates the clear sample size benefits of the GS designs compared to single analysis, fixed sample designs. The effect diminishes as variance increases due to the lowered power of the first test with decreasing noncentrality of the test statistic. When futility stopping is allowed, the expected sample size for the GS design decreases slightly as the probability of false futility stopping at the first stage analysis is introduced.

For an effect of twice the alternative of interest ($\theta = 2\theta_1$), the GS designs offer significant expected sample size reduction from the fixed design at all γ considered. The effect diminishes somewhat as first stage power decreases for increasing variance. There is virtually no difference in the two GS designs considered under this condition as the chance of first stage futility stopping is very small for $\theta = 2\theta_1$.

Table 4.8: $E(N_w)$ for Example 4.1: IPIA

γ	IPIA (no futility)			IPIA (futility)		
	$\theta = \mathbf{0}$	$\theta = \theta_1$	$\theta = 2\theta_1$	$\theta = \mathbf{0}$	$\theta = \theta_1$	$\theta = 2\theta_1$
0.50	42.7	40.2	39.0	42.0	40.1	39.0
0.75	58.4	50.0	39.2	55.1	50.0	39.2
1.00	77.1	67.8	42.3	70.8	67.5	42.3
1.50	115.1	109.4	71.9	102.8	107.7	71.9
2.00	153.2	150.2	122.1	134.9	146.1	122.0

Table 4.8 displays the values for expected sample size for the IPIA designs with and without futility stopping. Similar to the IP design, IPIA expected samples sizes are lower than the fixed design sample size for low true variance and increase with the true variance as the first stage sample size re-estimation requires larger second stage samples on average. Under the null hypothesis ($\theta = \mathbf{0}$), the IPIA design without futility stopping has similar sample size values to the IP design since stopping at the first stage for efficacy is rare here. The difference comes from the use of different distributions for sample size re-estimation (IP uses F , IPIA uses χ^2 here). The null case IPIA design with possible futility stopping causes a drop in expected sample size for all variance values when compared the design without futility stopping due to the chance of a correct decision to accept the null and stop at the first stage analysis.

Under the alternative of interest, ($\theta = \theta_1$), the expected sample sizes for the IPIA designs are noticeably lower than the fixed sample design for variance values at the preplanned value or lower. The expected sample sizes rise with γ due to the need for increased sample size detected at the first stage to protect study power. In this case, early

stopping (GS-like) sample size benefits are offset by the sample size recalculation procedure (power protecting IP characteristic). Expected sample size under the alternative of interest is slightly lower in the IPIA design allowing futility stopping due to the possibility of false futility stops at first stage. This chance increases with γ due to the naive p-value based futility stopping rules used to calculate the first stage futility critical value.

For an effect of twice the alternative of interest ($\theta = 2\theta_1$), the IPIA designs offer substantial expected sample size reduction due the large chance of early efficacy stopping in the first stage. The effect diminishes as increasing variance calls for more sample size in the second stage and decreases first stage power. There is very little difference in the two IPIA designs considered under $\theta = 2\theta_1$ as the chance of futility stopping is very small for the large effect size.

4.6 DISCUSSION

In this chapter, I have derived theory that generalizes the two-stage internal pilot with interim analysis for use for multiple degree of freedom univariate Gaussian linear model hypotheses. The exact results allow for accurate numerical calculation of type I error rate, power, and expected sample size for various study designs without the need for simulations. Many prospective research studies and even clinical trials are not large enough for asymptotic properties to hold. Since the theory in this chapter is not derived using asymptotic results, it will be accurate and valuable for planning smaller studies.

This results from Example 4.1 highlight some of the different characteristics of the designs considered, all of which are special cases of the IPIA design and theory detailed in this chapter. Each design has its advantages and disadvantages.

The fixed sample design has the advantages of a known sample size and a controlled type I error rate, but has an uncontrolled power function affected by the unknown true variance, a nuisance parameter. The GS design allows a study to stop early if effect size differs from the preplanned magnitude, and hence, decreases expected sample size from the

fixed sample level. In the tables above it achieved this goal under all conditions when futility stopping is allowed, and under situations of high effect size when futility stopping is not included. Using large sample (χ^2) critical values, the GS designs have an inflation in type I error rate due the critical values not accounting for uncertainty of the variance estimate. Finally, GS power is vulnerable to misspecification of variance at planning stage as shown in Table 4.6. This sensitivity is similar to that found in the fixed sample design and is due to the lack of sample size re-estimation for the final stage sample size.

The primary goal of the IP design is to protect study power by re-estimating sample size through interim power analysis without interim data analysis. As Table 4.6 shows, this goal is greatly achieved by the design. The IP design can also have sample size benefits due to possible sample size reduction at the interim stage if the planning variance value was specified higher than the true parameter value. For high true variance values, the design has higher expected sample sizes than the GS and fixed sample designs. The IP design also has inherent type I error rate inflation dependent on the true variance value. This is typically accounted for in smaller sample studies by adjusting the test statistic or critical value. Adjustment was not made here for comparative purposes.

The IPIA designs seek to incorporate the advantages of the GS and IP designs by allowing for early stopping as well as sample size re-estimation at the interim stage. Table 4.4 shows that it does, in fact, mostly achieve the power protective properties of the IP design. Also, Table 4.8 shows that the sample size benefits of the GS design are also inherent to the IPIA design as the expected sample sizes for many conditions are significantly lower than for the IP design.

It is apparent that the challenge of the IPIA design is controlling the type I error rate while maintaining the power and sample size benefits of the design. In addition to enjoying the benefits of power and sample size reduction from the IP and GS designs, respectively, it retains the different sources of potential type I error rate inflation that the designs introduce.

Careful adjustments that can control the type I error rate while maintaining the design's benefits must be done in order for the IPIA procedure to be useful in practice.

The IPIA procedure as outlined in this chapter is purposefully kept general in many regards. For example, it does not specify mandatory methods for selecting critical values, updating sample size, or selecting the interim stage sample size. The theory developed in this chapter as well as the further development of the prototype software to assist in calculation would allow for the exploration of many different possible designs. This would not only be valuable for the development of general study guidelines with positive characteristics. Also, since all studies are not alike, extensive exploration for a specific study during planning stages can allow investigators to customize the procedure for their specific needs.

Procedural strategies for single degree of freedom hypotheses within the IPIA framework were examined in Chapter 3 of this dissertation. I found that using the t distribution for critical value computation had a significant effect of reducing the type I error rate for the IPIA design. Also, I introduced a bounding method which controlled the type I error rate while maintaining the power and sample size benefits of the design. While I have not fully explored these methods within the more complex settings considered here, I believe that they have great promise. Unfortunately the computational intensity of the current IPIA bounding method would be burdensome to employ here due to the added layer of numerical integration of the results for this chapter. However, straightforward attention to choosing better numerical integration methods will likely radically speed the calculations.

I believe that using the F distribution for all critical value calculation is a good start towards further improvements. Developing and refining design strategies for the IPIA design, especially for complex hypotheses, will be a main topic of future research and is discussed in Chapter 5.

CHAPTER 5. SUMMARY AND FUTURE RESEARCH

5.1 SUMMARY OF ACCOMPLISHMENTS

5.1.1 Chapter 2: Internal Pilot with Interim Analysis for Single Degree of Freedom Hypothesis Tests

In Chapter 2 I 1) introduced the proposed model of a two-stage internal pilot with interim analysis (IPIA) design and 2) derived the exact distributional theory needed for planning studies with single degree of freedom hypothesis tests. The exact theory applies to any single degree of freedom hypothesis in a univariate linear model with fixed predictors, Gaussian errors, and unknown variance, including one and two group comparisons. Direct computation using the theory allows for fast calculation of power, type I error rate, and expected sample size. In the example considered, simulations took from 24-70 times more computational time. Also, I compared study characteristics of various designs and concluded that the IPIA was the only one able to control power at the level desired and simultaneously achieve sample size savings (when available) over a wide range of conditions. The numerical enumerations demonstrated the need for active intervention, above and beyond popular group sequential corrections, in order to guarantee control of the type I error rate.

5.1.2 Chapter 3: Planning Procedures for an Internal Pilot with Interim Analysis Design

In Chapter 3, I focused on design strategies for the IPIA for the single degree of freedom Gaussian linear models considered in Chapter 2. The goal was to achieve sound study design strategies that control the type I error rate while best maintaining the power and sample size advantages of the design. I introduced the IPIA bounding method which allows for type I error controlled IPIA designs. I showed the importance of accounting for the

uncertainty of the variance estimate in test statistics through use of t distribution based critical value calculation. Using the t with the bounding method gave not only good control of type I error rate for the IPIA design, but also gave good control of power and sample size, even in small sample studies. Finally, I demonstrated and briefly explored the complexity of study properties with respect to the interim sample size decision.

5.1.3 Chapter 4: Internal Pilot with Interim Analysis for Multiple Degree of Freedom Hypothesis Tests

Chapter 4 also centered on the model and IPIA design introduced in Chapter 2, but generalized the results to *any* general linear hypothesis with one or more degrees of freedom. I introduced new exact theory that allows for accurate study planning for complex designs even within small sample studies. The theory includes an explicit and computable form for the conditional joint distribution of the first and second stage test statistics. Together with the results from Chapter 2, the distribution allows calculating power, type I error rate, and expected sample size for the models. Through examples, I compared study characteristics of an IPIA design with the characteristics of a fixed sample design, internal pilot, and a two-stage group sequential design, all of which are special cases within the IPIA framework. In the numerical enumerations considered, the IPIA design best protected power while allowing for sample size saving (when available) over a range of conditions. The levels of type I error rate inflation observed in the examples demonstrated the need for design and analysis methods more complex than traditional group-sequential corrections based on large sample theory.

5.2 FUTURE RESEARCH

5.2.1 Futility Bounds

The theory developed in Chapters 2 and 3 optionally allows for stopping under the null at the interim analysis if $F_1 < f_l$, where f_l is the first stage lower critical value. This can cause a great reduction in expected sample size when the effect size is near the null value by allowing the study to stop for a "lost cause". If no early futility stopping ability is desired, then the lower critical value, f_l , is set to zero. In the examples in Chapters 2 and 4, a simple p-value based futility bound was employed: a study was stopped for futility when the p-value for the first stage test was greater than 0.85. This simple method clearly showed the possible resource savings of a carefully used futility bound procedure. In all cases when no effect was present, there was a noticeable drop in expected sample size due to early stopping under the null. I believe that further development in this area for the IPIA model is extremely important.

Current methods exist for futility stopping within group sequential methods (Pampallona and Tsiatis, 1994; Lachin, 2005). Adapting these methods for use in the IPIA setting could be a good start into an IPIA futility analysis plan. The IPIA is simpler than the group sequential design in that it only has two stages, but more complex in that sample size is unknown and based on a random variance estimate. Accuracy in small samples requires accounting for the uncertainty of the variance at the interim stage. One possible method would be to solve futility bounds based on conditional power calculations assuming variance to be a lower confidence limit (optimistic) at the interim stage. This would assure that a study only stops when the probability of a significant outcome if continued is low.

5.2.2 Sample Size Re-Estimation Method

For continuous Gaussian data, sample size re-estimation is an interim power analysis that uses the variance estimate from the first stage to calculate the estimated sample size needs of a study. The sample size re-estimation rule determines the distribution of N_+ by

taking advantage of the monotone relationship between continuous $\hat{\sigma}_1^2$ and discrete N_+ . It is an important consideration in the IPIA design affecting type I error rate, power, and expected sample size.

For general linear hypotheses within the IPIA framework, this process is complicated by the lack of knowledge of the final test statistic distribution as well as by the difficulty in knowing the correct critical value to use. The final test statistic is not a true F distributed variable. Both the choice of continuation and selection of sample size complicate and bias the effect size and variance estimates used in the test statistic calculation.

The critical value selection at interim power analysis is also complex. As is typical in group sequential sample size calculation, I used the target type I error rate to assume a final critical value for sample size re-estimation purposes. In reality, the amount of type I error spent on the second stage is not equal to the target level due to early testing. The issue is complicated by the feedback between critical value and sample size. The critical value determines the sample size needs and the sample fraction determines the critical value to use. Research into better aligning the true test statistic distribution and critical values used with the assumptions at the interim power analysis seems likely to boost IPIA study efficiency and accuracy.

5.2.3 Selection of Interim Sample Size

The value of the interim sample size, n_1 , is an important consideration in determining the performance characteristics of an IP or IPIA design. In Chapter 3, I illustrated the complexity engendered by the choice of n_1 for the IPIA model. Depending on the critical values employed, true effect size, and ratio of planning and true variance values, a high value of n_1 could possibly have sample size savings due to the changing power at the first stage test. Because of the interactive nature and sensitivity of type I error, power, and expected sample size to the interim sample size, the effects should be explored in some detail as part of study planning.

5.2.4 Computation

New methods will rarely be adopted without a convenient means to use them. Consequently I feel it is extremely important to produce accurate and user-friendly software for the study planning and data analysis with an IPIA design. The open-source software would also facilitate future research using these results as a foundation.

During the course of this research, a large amount of code was written for the calculations completed. The code consists of a number of SAS/IML (SAS Institute, 2004) modules that together form a prototype program. The code supports calculation of type I error rate, power (under any alternative), and expected sample size for any general linear univariate model with Gaussian errors and fixed predictors. Additionally, the code allows using the IPIA bounding method for the single degree of freedom hypotheses as described in Chapter 3. The code can calculate the location and value of maximum type I error rate as a function of true variance, or it can find the adjusted rate with bounded type I error rate inflation (see Chapter 3). For single degree of freedom IPIA tests, the current code works many times faster than simulation. While the developed code has worked well for the needs at hand, it is a prototype in that little effort has been made at increasing its usability, error-checking of inputs, and efficiency of calculation. Further development in this area would not only allow the IPIA methods to more easily be adopted, but would facilitate my personal research and that of others by making an efficient foundation for future development.

5.2.5 Strategies for Multiple Degree of Freedom Tests

In Chapter 3 I examined IPIA design strategies for single degree of freedom hypothesis tests. I believe that many of the recommended methods would also be applicable in the multiple degree of freedom setting. For example, the use of the F distribution instead of the large sample χ^2 distribution would more accurately model the distribution of the test statistic. Obviously the bounding method will work within the more complex setting, with computational speed the only current barrier.

5.2.6 Generalizations to Other Settings

Multiple Stage Designs

In this research, I considered only two-stage designs. The priority was to combine the power protection of the internal pilot design with the ability to save resources through early stopping. Traditional group sequential methods allow for more than two looks during the course of a study. It would be valuable to theory and methods to allow for a larger number of looks during the course of a study. I feel that a good approximation may be achieved in two simple steps. First, employ the current IPIA bounding method for a two-stage design to choose an adjusted overall alpha as a first step. Second, traditional group sequential rules would be used to split alpha across stages.

Multivariate and Repeated Measures Models

The current theory was developed within the general linear univariate model framework. Generalizing it to studies with multivariate and repeated measures models would greatly increase its range of usefulness. The general linear multivariate model may be stated as $Y = XB + E$, with $\{Y, B, E\}$ having p columns corresponding to p responses. Interest lies in testing the null hypothesis $H_0 : \Theta = \Theta_0$ for $\Theta = CBU$ of dimensions $a \times b$.

A particularly interesting class of multivariate hypotheses, $a = 1$ and $b > 1$, has recently been shown by Park (2007) to have equivalent univariate forms. She was therefore also able to show that such models fit exactly into the framework and exact theory of the univariate internal pilot. Interesting applications include any one or two group comparison of multivariate responses or profiles (such as time trends). Hotelling one and two sample tests are special cases. I believe the IPIA theory can also be applied exactly to this particular multivariate setting.

In the general case of $\min(a, b) > 1$, it helps to distinguish between the MULTIREP tests based on the affine invariant statistics used for the multivariate approach to repeated

measures, and the UNIREP tests based on the orthonormal invariant test used for the univariate approach to repeated measures (Muller and Stewart, 2006, Chapter 3, et seq.). Park (2007) also indicated that the MULTIREP approximate methods reviewed in Muller et al. (1992) could in a parallel way give equally accurate approximations for $\min(a, b) > 1$. Similarly, general approximations for the UNIREP case (Muller et al., 2007) as used in Coffey and Muller (2003) for internal pilots, should provide equally accurate approximations for IPIA designs.

Mixed Models

The proposals for future models to consider reflect the 'Divide and Conquer' approach recommended for mixed models by Gurka, Coffey, and Muller in an invited presentation at the Joint Statistical Meetings of 2007. Following the approach of piecemeal adaptation, Gurka et al. (2007) described a useful class of mixed models that can be expressed as equivalent univariate tests and work exactly within developed internal pilot theory. This class includes complete and balanced designs with compound symmetric covariance. Detailing how the IPIA theory can be used with this class of designs models would bring further generalization to the methods. In turn, Johnson's work (2007) on cluster samples can be understood as generalizing the exact results of Gurka et al. to unbalanced designs. Her work therefore indicates how to extend the IPIA model in a parallel way.

APPENDIX A: CHAPTER 2 PROOFS

Lemma A.1 Any testable general linear hypothesis with $\boldsymbol{\theta}_0 \neq \mathbf{0}$ may be expressed in terms of a related model and general linear hypothesis with $\boldsymbol{\theta}_0 = \mathbf{0}$.

Proof. Transforming the model gives

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} && + \mathbf{e} \\ \mathbf{y} - \mathbf{X}\mathbf{C}'(\mathbf{C}\mathbf{C}')^{-1}\boldsymbol{\theta}_0 &= \mathbf{X}[\boldsymbol{\beta} - \mathbf{C}'(\mathbf{C}\mathbf{C}')^{-1}\boldsymbol{\theta}_0] && + \mathbf{e} \\ \mathbf{y}_* &= \mathbf{X}\boldsymbol{\beta}_* && + \mathbf{e} . \end{aligned} \tag{A.1}$$

Hence $\boldsymbol{\theta}_* = \mathbf{C}\boldsymbol{\beta}_* = \boldsymbol{\theta} - \boldsymbol{\theta}_0$ and $H_0 : \boldsymbol{\theta}_* = \mathbf{0}$ is equivalent to $H_0 : \mathbf{C}\boldsymbol{\beta} = \boldsymbol{\theta}_0$. □

Corollary 2.1 Proof. From equation 2.58, for $a = 1$, $\boldsymbol{\mu}_{e.h} = (n_+/n_2)t_0\mathbf{V}'_{ep}\mathbf{v}_{h1}$ and $\boldsymbol{\Sigma}_{e.h} = \sigma^2[\mathbf{I}_{n_2} - (n_+/n_2)\mathbf{V}'_{ep}\mathbf{A}_{h1}\mathbf{V}_{ep}]$ the following distribution hold:

$$\mathbf{y}_{e.h} = \mathbf{y}_{ep} | (\mathbf{v}'_{h1}\mathbf{V}_{ep}\mathbf{y}_{ep} = t_0) \sim (S)\mathcal{N}_{n_2}(\boldsymbol{\mu}_{e.h}, \boldsymbol{\Sigma}_{e.h}) .$$

So, since $E_p = \mathbf{y}'_{ep}\mathbf{y}_{ep}/\sigma^2$, the variable $E_p | (\mathbf{v}'_{h1}\mathbf{V}_{ep}\mathbf{y}_{ep} = t_0)$ is distributed the same as

$$q = \mathbf{y}'_{e.h}\mathbf{y}_{e.h}/\sigma^2 = \mathbf{y}'_{e.h}\mathbf{A}\mathbf{y}_{e.h} \tag{A.2}$$

for $\mathbf{A} = \mathbf{A}' = (1/\sigma^2)\mathbf{I}_{n_2}$. Also, for $\mathbf{F}_A = \sigma^{-1}\mathbf{I}_{n_2}$, \mathbf{A} can be written

$$\mathbf{A} = \mathbf{F}_A\mathbf{F}'_A . \tag{A.3}$$

For $\boldsymbol{\mu}_A = \mathbf{F}'_A\boldsymbol{\mu}_{e.h} = \sigma^{-1}\boldsymbol{\mu}_{e.h}$ and $\boldsymbol{\Sigma}_A = \mathbf{F}'_A\boldsymbol{\Sigma}_{e.h}\mathbf{F}_A = (1/\sigma^2)\boldsymbol{\Sigma}_{e.h}$, define

$$\mathbf{y}_A = \mathbf{F}'_A\mathbf{y}_{e.h} \sim (S)\mathcal{N}_{n_2}(\boldsymbol{\mu}_A, \boldsymbol{\Sigma}_A) . \tag{A.4}$$

Also note,

$$\begin{aligned} \boldsymbol{\Sigma}_A &= \mathbf{I}_{n_2} - (n_+/n_2)\mathbf{V}'_{ep}\mathbf{A}_{h1}\mathbf{V}_{ep} \\ &= \mathbf{I}_{n_2} - (n_+/n_2)(\mathbf{V}'_{ep}\mathbf{v}_{h1})(\mathbf{v}'_{h1}\mathbf{V}_{ep}) \\ &= \mathbf{I}_{n_2} - (n_+/n_2)(\mathbf{v}_{x1})(\mathbf{v}'_{x1}) \\ &= \mathbf{I}_{n_2} - (n_+/n_2)\left\{ \left[(\mathbf{V}'_{ep}\mathbf{v}_{h1})(n_2/n_+)^{-1/2} \right] (n_2/n_+) \left[(n_2/n_+)^{-1/2}\mathbf{v}'_{h1}\mathbf{V}_{ep} \right] \right\} \\ &= \mathbf{I}_{n_2} - (n_+/n_2)\mathbf{v}_{x2}(n_2/n_+)\mathbf{v}'_{x2} \\ &= \mathbf{I}_{n_2} - \mathbf{v}_{x2}\mathbf{v}'_{x2} \end{aligned} \tag{A.5}$$

Here $(\mathbf{v}_{x1}\mathbf{v}'_{x1})$ has rank 1, with only nonzero eigenvalue $(\mathbf{v}'_{h1}\mathbf{V}_{ep})(\mathbf{V}'_{ep}\mathbf{v}_{h1}) = (n_2/n_+)$ because, in general, $\mathbf{X}\mathbf{X}'$ and $\mathbf{X}'\mathbf{X}$ have same nonzero eigenvalues and the eigenvalue of a scalar the number itself. Hence $\Sigma_A = \mathbf{I}_{n_2} - (n_+/n_2)\mathbf{V}'_{ep}\mathbf{A}_{h1}\mathbf{V}_{ep}$ is $n_2 \times n_2$, rank $(n_2 - 1)$, and idempotent, so has $(n_2 - 1)$ eigenvalues of one and one of zero. Hence, Σ_A has rank $n_2 - 1$.

By spectral decomposition one can write

$$\Sigma_A = [\Psi \quad \mathbf{v}_{x2}] \begin{bmatrix} \mathbf{I}_{n_2-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \Psi' \\ \mathbf{v}'_{x2} \end{bmatrix} . \quad (\text{A.6})$$

Then

$$\begin{aligned} \Psi^+ &= (\Psi'\Psi)^{-1}\Psi' \\ &= \Psi' . \end{aligned} \quad (\text{A.7})$$

Also,

$$[\Psi \quad \mathbf{v}_{x2}] \begin{bmatrix} \Psi' \\ \mathbf{v}'_{x2} \end{bmatrix} = \Psi\Psi' + \mathbf{v}_{x2}\mathbf{v}'_{x2} = \mathbf{I}_{n_2} . \quad (\text{A.8})$$

So

$$\mu_A = \Psi\Psi'\mu_A + \mathbf{v}_{x2}\mathbf{v}'_{x2}\mu_A . \quad (\text{A.9})$$

Define $\mu_z = \Psi'\mu_A$ and $\mu_0 = \mathbf{v}_{x2}\mathbf{v}'_{x2}\mu_A$. If $\mathbf{y}_0 = \mathbf{y}_A - \mu_A$ and $\mathbf{z} \sim \mathcal{N}_{n_2}(\mathbf{0}, \mathbf{I}_{n_2})$, then $\mathbf{y}_0 \sim (\mathcal{S})\mathcal{N}_{n_2}(\mathbf{0}, \Sigma_A)$ and

$$\begin{aligned} \mathbf{y}_A &= \mathbf{y}_0 + \mu_A \\ &= \mathbf{y}_0 + \Psi\Psi'\mu_A + \mathbf{v}_{x2}\mathbf{v}'_{x2}\mu_A \\ &= \Psi\mathbf{z} + \Psi\Psi'\mu_A + \mu_0 \\ &= \Psi[\mathbf{z} + \Psi'\mu_A] + \mu_0 \\ &= \Psi(\mathbf{z} + \mu_z) + \mu_0 . \end{aligned}$$

(A.10)

Now the following can be expressed:

$$\begin{aligned}
q &= \mathbf{y}'_A \mathbf{y}_A \\
&= [\mathbf{\Psi}(\mathbf{z} + \boldsymbol{\mu}_z) + \boldsymbol{\mu}_0]' [\mathbf{\Psi}(\mathbf{z} + \boldsymbol{\mu}_z) + \boldsymbol{\mu}_0] \\
&= (\mathbf{z} + \boldsymbol{\mu}_z)' \mathbf{\Psi}' \mathbf{\Psi} (\mathbf{z} + \boldsymbol{\mu}_z) + 2(\mathbf{z} + \boldsymbol{\mu}_z)' \mathbf{\Psi}' \boldsymbol{\mu}_0 + \boldsymbol{\mu}'_0 \boldsymbol{\mu}_0 \\
&= (\mathbf{z} + \boldsymbol{\mu}_z)' (\mathbf{z} + \boldsymbol{\mu}_z) + 2(\mathbf{z} + \boldsymbol{\mu}_z)' \mathbf{\Psi}' \mathbf{v}_{x2} \mathbf{v}'_{x2} \boldsymbol{\mu}_A + \boldsymbol{\mu}'_0 \boldsymbol{\mu}_0 \\
&= \sum_{k=1}^{n_2-1} (z_k + \mu_{z(k)})^2 + \boldsymbol{\mu}'_0 \boldsymbol{\mu}_0 .
\end{aligned} \tag{A.11}$$

Also, since

$$\begin{aligned}
\boldsymbol{\mu}_z &= \mathbf{\Psi}^+ \boldsymbol{\mu}_A \\
&= \sigma^{-1} \mathbf{\Psi}' \boldsymbol{\mu}_{e,h} \\
&= \sigma^{-1} \mathbf{\Psi}' (\mathbf{V}'_{ep} \mathbf{v}_{h1}) \\
&= \sigma^{-1} \mathbf{\Psi}' \mathbf{v}_{x2} \\
&= \mathbf{0}
\end{aligned} \tag{A.12}$$

and

$$\begin{aligned}
\boldsymbol{\mu}'_0 \boldsymbol{\mu}_0 &= \boldsymbol{\mu}'_A \mathbf{v}_{x2} \mathbf{v}'_{x2} \mathbf{v}_{x2} \mathbf{v}'_{x2} \boldsymbol{\mu}_A \\
&= \boldsymbol{\mu}'_A \mathbf{v}_{x2} \mathbf{v}'_{x2} \boldsymbol{\mu}_A \\
&= \boldsymbol{\mu}'_A (\mathbf{I} - \boldsymbol{\Sigma}_A) \boldsymbol{\mu}_A \\
&= \sigma^{-2} \boldsymbol{\mu}'_{e,h} (\mathbf{I} - \sigma^{-2} \boldsymbol{\Sigma}_A) \boldsymbol{\mu}_{e,h} \\
&= (n_+/n_2)^2 (t_0/\sigma)^2 \mathbf{V}'_{h1} \mathbf{V}_{ep} [(n_+/n_2) \mathbf{V}'_{ep} \mathbf{A}_{h1} \mathbf{V}_{ep}] \mathbf{V}'_{ep} \mathbf{V}_{h1} \\
&= (n_+/n_2)^3 (t_0/\sigma)^2 (\mathbf{V}'_{h1} \mathbf{A}_{ep} \mathbf{V}_{h1}) (\mathbf{V}'_{h1} \mathbf{A}_{ep} \mathbf{V}_{h1}) \\
&= (n_+/n_2)^3 (t_0/\sigma)^2 (n_2/n_+) (n_2/n_+) \\
&= (n_+/n_2) (t_0/\sigma)^2 ,
\end{aligned} \tag{A.13}$$

one can write

$$q = \sum_{k=1}^{n_2-1} z_k^2 + (n_+/n_2) (t_0/\sigma)^2 = X_p + (n_+/n_2) (t_0/\sigma)^2 \tag{A.14}$$

where X_p is a central χ^2 distributed variable with $n_2 - 1$ degrees of freedom.

□

Theorem 2.1 Proof. Related results and definitions can be found in section 2.2.3. For

$c_s = f_s/\nu_s$, $E_s = \nu_s \widehat{\sigma}_s^2/\sigma^2$, $E_p = E_+ - E_1$, $G_s = \widehat{\delta}_s/\sigma^2$ with $s \in \{1, +\}$. Define
 $b(t_e, t_h) = t_e^2/(\sigma^2 c_+) - t_e$, $d(t_p) = n_+ t_p^2/(n_2 \sigma^2)$, $h(t_e, t_h) = \sigma(c_1 t_e)^{1/2} - (n_1/n_+)^{1/2} t_h$,
 $l(t_e, t_h) = -\sigma(c_1 t_e)^{1/2} - (n_1/n_+)^{1/2} t_h$, and $p_{n_+}^{-1} = F_{\chi^2}(q_2; \nu_1) - F_{\chi^2}(q_1; \nu_1)$.

$$\begin{aligned}
& F_{F_1, F_+ | N_+}(f_1, f_+) \\
&= \Pr_{N_+} \{F_1 \leq f_1, F_+ \leq f_+\} \\
&= \Pr_{N_+} \left\{ \widehat{\delta}_1 / \widehat{\sigma}_1^2 \leq f_1, \widehat{\delta}_+ / \widehat{\sigma}_+^2 \leq f_+ \right\} \\
&= \Pr_{N_+} \{G_1 \leq c_1 E_1, G_+ \leq c_+(E_1 + E_p)\} \\
&= p_{n_+}^{-1} \int_{q_1}^{q_2} f_{\chi^2}(t_e; \nu_1) \Pr_{N_+} \{G_1 \leq c_1 t_e, G_+ \leq c_+(t_e + E_p)\} dt_e \\
&= p_{n_+}^{-1} \int_{q_1}^{q_2} f_{\chi^2}(t_e; \nu_1) \\
&\quad \times \Pr_{N_+} \left\{ \left| \mathbf{v}'_{h1} \mathbf{V}_{ep} \mathbf{y}_{ep} + \sqrt{\frac{n_1}{n_+}} y_{h+} \right| \leq \sigma \sqrt{c_1 t_e}, y_{h+}^2 \leq \sigma^2 c_+(t_e + E_p) \right\} dt_e \\
&= p_{n_+}^{-1} \int_{q_1}^{q_2} f_{\chi^2}(t_e; \nu_1) \int_{-\infty}^{\infty} f_{\mathcal{N}}(t_h; \mu_{h+}, \sigma^2) \\
&\quad \times \Pr_{N_+} \{l(t_e, t_h) \leq \mathbf{v}'_{h1} \mathbf{V}_{ep} \mathbf{y}_{ep} \leq h(t_e, t_h), b(t_e, t_h) \leq E_p\} dt_h dt_e \\
&= p_{n_+}^{-1} \int_{q_1}^{q_2} f_{\chi^2}(t_e; \nu_1) \int_{-\infty}^{\infty} f_{\mathcal{N}}(t_h; \mu_{h+}, \sigma^2) \int_{l(t_e, t_h)}^{h(t_e, t_h)} f_{\mathcal{N}}[t_p; \mathbf{0}, (n_2/n_+) \sigma^2] \\
&\quad \times \Pr_{N_+} \{b(t_e, t_h) \leq E_p | \mathbf{v}'_{h1} \mathbf{V}_{ep} \mathbf{y}_{ep}\} dt_p dt_h dt_e \\
&= p_{n_+}^{-1} \int_{q_1}^{q_2} f_{\chi^2}(t_e; \nu_1) \int_{-\infty}^{\infty} f_{\mathcal{N}}(t_h; \mu_{h+}, \sigma^2) \int_{l(t_e, t_h)}^{h(t_e, t_h)} f_{\mathcal{N}}[t_p; \mathbf{0}, (n_2/n_+) \sigma^2] \\
&\quad \times \Pr_{N_+} \{b(t_e, t_h) - d(t_p) \leq X_p\} dt_p dt_h dt_e \\
&= p_{n_+}^{-1} \int_{q_1}^{q_2} f_{\chi^2}(t_e; \nu_1) \int_{-\infty}^{\infty} f_{\mathcal{N}}(t_h; \mu_{h+}, \sigma^2) \int_{l(t_e, t_h)}^{h(t_e, t_h)} f_{\mathcal{N}}[t_p; \mathbf{0}, (n_2/n_+) \sigma^2] \\
&\quad \times \{1 - F_{\chi^2}[b(t_e, t_h) - d(t_p); n_2 - 1]\} dt_p dt_h dt_e
\end{aligned}$$

□

Theorem 2.2 Proof. Related results and definitions can be found in section 2.2.3. For $n_2 = 1$, $c_s = f_s/\nu_s$, $E_s = \nu_s \hat{\sigma}_s^2/\sigma^2$, $E_p = E_+ - E_1$, $G_s = \hat{\delta}_s/\sigma^2$ with $s \in \{1, +\}$. Define $b(t_e) = n_+ c_1 t_e$ and $p_{n_+}^{-1} = F_{\chi^2}(q_2; \nu_1) - F_{\chi^2}(q_1; \nu_1)$.

$$\begin{aligned}
& F_{F_1, F_+ | N_+}(f_1, f_+) \\
&= \Pr_{N_+} \{F_1 \leq f_1, F_+ \leq f_+\} \\
&= \Pr_{N_+} \left\{ \hat{\delta}_1 / \hat{\sigma}_1^2 \leq f_1, \hat{\delta}_+ / a \cdot \hat{\sigma}_+^2 \leq f_+ \right\} \\
&= \Pr_{N_+} \{G_1 \leq c_1 E_1, G_+ \leq c_+(E_1 + E_p)\} \\
&= p_{n_+}^{-1} \int_{q_1}^{q_2} f_{\chi^2}(t_e; \nu_1) \Pr_{N_+} \{G_1 \leq c_1 t_e, G_+ \leq c_+(t_e + E_p)\} dt_e \\
&= p_{n_+}^{-1} \int_{q_1}^{q_2} f_{\chi^2}(t_e; \nu_1) \\
&\quad \times \Pr_{N_+} \left\{ \left(\sqrt{\frac{n_2}{n_+}} y_{ep} + \sqrt{\frac{n_1}{n_+}} y_{h+} \right)^2 \leq \sigma^2 c_1 t_e, G_+ \leq c_+(t_e + E_p) \right\} dt_e \\
&= p_{n_+}^{-1} \int_{q_1}^{q_2} f_{\chi^2}(t_e; \nu_1) \\
&\quad \times \Pr_{N_+} \{(E_p + n_1 G_+) \leq n_+ c_1 t_e, G_+ \leq c_+(t_e + E_p)\} dt_e \\
&= p_{n_+}^{-1} \int_{q_1}^{q_2} f_{\chi^2}(t_e; \nu_1) \int_0^{b(t_e)} f_{\chi^2}(t_p; n_2) \\
&\quad \times \Pr_{N_+} \{G_+ \leq [b(t_e) - t_p]/n_1, G_+ \leq c_+(t_e + t_p)\} dt_p dt_e \\
&= p_{n_+}^{-1} \int_{q_1}^{q_2} f_{\chi^2}(t_e; \nu_1) \int_0^{b(t_e)} f_{\chi^2}(t_p; n_2) \\
&\quad \times F_{\chi^2} \{ \min[(b(t_e) - t_p)/n_1, c_+(t_e + t_p)], 1, \lambda_+ \} dt_h dt_e
\end{aligned}$$

□

Theorem 2.3 Proof.

$$\begin{aligned}
F_{F_1|N_+}(f_1) &= \Pr\{F_1 \leq f_1 | N_+ = n_+\} \\
&= \Pr\left\{\widehat{\delta}_1 / (a \cdot \widehat{\sigma}_1^2) \leq f_1 | N_+ = n_+\right\} \\
&= \Pr\{G_1 \leq c_1 E_1 | N_+ = n_+\} \\
&= \int_{q_1}^{q_2} \frac{\Pr(G_1 \leq c_1 t_1) f_{\chi^2}(t_1; \nu_1)}{F_{\chi^2}(q_2; \nu_1) - F_{\chi^2}(q_1; \nu_1)} dt_1 \\
&= \int_{q_1}^{q_2} \frac{F_{\chi^2}(c_1 t_1; a, \lambda_1) f_{\chi^2}(t_1; \nu_1)}{F_{\chi^2}(q_2; \nu_1) - F_{\chi^2}(q_1; \nu_1)} dt_1
\end{aligned}$$

□

Corollary 2.2 Proof.

$$\begin{aligned}
&\Pr\{f_l \leq F_1 \leq f_u | N_+ = n_+\} \\
&= \Pr\left\{f_l \leq \frac{\widehat{\delta}_1 / a}{\widehat{\sigma}_1^2} \leq f_u | N_+ = n_+\right\} \\
&= \Pr\{c_l E_1 \leq G_1 \leq c_u E_1 | N_+ = n_+\} \\
&= \int_{q_1}^{q_2} \frac{\Pr\{c_l t_1 \leq G_1 \leq c_u t_1\} f_{\chi^2}(t_1; \nu_1)}{\Pr\{N_+ = n_+\}} dt_1 \\
&= \int_{q_1}^{q_2} \frac{[F_{\chi^2}(c_u t_1; a, \lambda_1) - F_{\chi^2}(c_l t_1; a, \lambda_1)] f_{\chi^2}(t_1; \nu_1)}{F_{\chi^2}(q_2; \nu_1) - F_{\chi^2}(q_1; \nu_1)} dt_1
\end{aligned}$$

□

Theorem 2.4 Proof. An expression for unconditional power, P_w , can be written

$$\begin{aligned}
P_w &= \Pr\{\text{Reject } H_0\} \\
&= 1 - \Pr\{\text{Accept } H_0\} \\
&= 1 - \Pr\{\text{Accept } H_0 \text{ at stage 1} \cup (\text{Continue at stage 1} \cap \text{Accept } H_0 \text{ at stage 2})\} \\
&= 1 - \sum_{\{N_+=n_+\}} \Pr(N_+ = n_+) \left\{ F_{F_1|N_+}[f_l(n_+)] + \right. \\
&\quad \left. \Pr\left[f_l(n_+) \leq F_1 < f_u(n_+), F_+ < f_+(n_+) | N_+ = n_+ \right] \right\}
\end{aligned}$$

$\Pr(N_+ = n_+) = [F_{\chi^2}(q_2; \nu_1) - F_{\chi^2}(q_1; \nu_1)]$ finishes the proof.

□

Theorem 2.5 Proof. For N_w defined as

$$N_w = \begin{cases} n_1 & \text{if study stopped after first stage} \\ N_+ & \text{otherwise} \end{cases}$$

the following holds:

$$\begin{aligned} E(N_w) &= \sum_{\{N_+=n_+\}} \left\{ n_1 [1 - \Pr(f_l \leq F_1 < f_u, N_+ = n_+)] \right. \\ &\quad \left. + n_+ \Pr(f_l \leq F_1 < f_u, N_+ = n_+) \right\} \\ &= n_1 + \sum_{\{N_+=n_+\}} n_2 \Pr(N_+ = n_+) \Pr[f_l \leq F_1 < f_u | N_+ = n_+] \\ &= n_1 + \sum_{\{N_+=n_+\}} n_2 [F_{\chi^2}(q_2; \nu_1) - F_{\chi^2}(q_1; \nu_1)] \Pr[f_l \leq F_1 < f_u | N_+ = n_+] \end{aligned}$$

□

APPENDIX B: CHAPTER 4 PROOFS

Lemma 4.1 Proof.

$$\begin{aligned}
 BX'_s X_s B &= (X'_0 X_0)^{-1} C' M_0^{-1} C (X'_0 X_0)^{-1} X'_s X_s (X'_0 X_0)^{-1} C' M_0^{-1} C (X'_0 X_0)^{-1} \\
 &= k_s (X'_0 X_0)^{-1} C' M_0^{-1} C (X'_0 X_0)^{-1} C' M_0^{-1} C (X'_0 X_0)^{-1} \\
 &= k_s (X'_0 X_0)^{-1} C' M_0^{-1} M_0 M_0^{-1} C (X'_0 X_0)^{-1} \\
 &= k_s (X'_0 X_0)^{-1} C' M_0^{-1} C (X'_0 X_0)^{-1} \\
 &= k_s B
 \end{aligned}$$

□

Lemma 4.2 Proof. a. Symmetry is clear from the expressed forms.

b. The results follow directly by using two simple results from Chapter 2, namely

$$X'_s X_s = k_s X'_0 X_0 \text{ and } M_s = k_s^{-1} M_0.$$

c. i.

$$\begin{aligned}
 A_{h+} A_{h1} &= k_+^{-1} k_1^{-1} \begin{bmatrix} X_1 B X'_1 & X_1 B X'_2 \\ X_2 B X'_1 & X_2 B X'_2 \end{bmatrix} \begin{bmatrix} X_1 B X'_1 & 0 \\ 0 & 0 \end{bmatrix} \\
 &= k_+^{-1} k_1^{-1} \begin{bmatrix} X_1 B X'_1 X_1 B X'_1 & 0 \\ X_2 B X'_1 X_1 B X'_1 & 0 \end{bmatrix} \\
 &= k_+^{-1} \begin{bmatrix} X_1 B X'_1 & 0 \\ X_2 B X'_1 & 0 \end{bmatrix}
 \end{aligned}$$

ii.

$$\begin{aligned}
 A_{h+} A_{h2} &= k_+^{-1} k_2^{-1} \begin{bmatrix} X_1 B X'_1 & X_1 B X'_2 \\ X_2 B X'_1 & X_2 B X'_2 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & X_2 B X'_2 \end{bmatrix} \\
 &= k_+^{-1} k_2^{-1} \begin{bmatrix} 0 & X_1 B X'_2 X_2 B X'_2 \\ 0 & X_2 B X'_2 X_2 B X'_2 \end{bmatrix} \\
 &= k_+^{-1} \begin{bmatrix} 0 & X_1 B X'_2 \\ 0 & X_2 B X'_2 \end{bmatrix}
 \end{aligned}$$

iii. Summing the results from parts i. and ii. gives

$$A_{h+} A_{h1} + A_{h+} A_{h2} = k_+^{-1} \begin{bmatrix} X_1 B X'_1 & X_1 B X'_2 \\ X_2 B X'_1 & X_2 B X'_2 \end{bmatrix} = A_{h+}$$

iv.

$$\begin{aligned}
\mathbf{A}_{h_+}\mathbf{A}_{h_1} + \mathbf{A}_{h_1}\mathbf{A}_{h_+} &= k_+^{-1} \begin{bmatrix} \mathbf{X}_1\mathbf{B}\mathbf{X}'_1 & \mathbf{0} \\ \mathbf{X}_2\mathbf{B}\mathbf{X}'_1 & \mathbf{0} \end{bmatrix} + k_+^{-1} \begin{bmatrix} \mathbf{X}_1\mathbf{B}\mathbf{X}'_1 & \mathbf{X}_1\mathbf{B}\mathbf{X}'_2 \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \\
&= k_+^{-1} \begin{bmatrix} 2\mathbf{X}_1\mathbf{B}\mathbf{X}'_1 & \mathbf{X}_1\mathbf{B}\mathbf{X}'_2 \\ \mathbf{X}_2\mathbf{B}\mathbf{X}'_1 & \mathbf{0} \end{bmatrix} \\
&= \mathbf{A}_{h_+} + k_+^{-1}k_1\mathbf{A}_{h_1} - k_+^{-1}k_2\mathbf{A}_{h_2}
\end{aligned}$$

v.

$$\begin{aligned}
\mathbf{A}_{h_+}\mathbf{A}_{h_2} + \mathbf{A}_{h_2}\mathbf{A}_{h_+} &= k_+^{-1} \begin{bmatrix} \mathbf{0} & \mathbf{X}_1\mathbf{B}\mathbf{X}'_2 \\ \mathbf{0} & \mathbf{X}_2\mathbf{B}\mathbf{X}'_2 \end{bmatrix} + k_+^{-1} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{X}_2\mathbf{B}\mathbf{X}'_1 & \mathbf{X}_2\mathbf{B}\mathbf{X}'_2 \end{bmatrix} \\
&= k_+^{-1} \begin{bmatrix} \mathbf{0} & \mathbf{X}_1\mathbf{B}\mathbf{X}'_2 \\ \mathbf{X}_2\mathbf{B}\mathbf{X}'_1 & 2\mathbf{X}_2\mathbf{B}\mathbf{X}'_2 \end{bmatrix} \\
&= \mathbf{A}_{h_+} - k_+^{-1}k_1\mathbf{A}_{h_1} + k_+^{-1}k_2\mathbf{A}_{h_2}
\end{aligned}$$

vi.

$$\begin{aligned}
\mathbf{A}_{e_+}\mathbf{A}_{e_1} &= (\mathbf{I}_{n_+} - \mathbf{H}_+)\mathbf{A}_{e_1} \\
&= \mathbf{A}_{e_1} - \mathbf{X}_+(\mathbf{X}'_+\mathbf{X}_+)^{-1}[\mathbf{X}'_1 \quad \mathbf{X}'_2] \begin{bmatrix} \mathbf{I}_{n_1} - \mathbf{H}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0}_{n_2 \times n_2} \end{bmatrix} \\
&= \mathbf{A}_{e_1} - \mathbf{X}_+(\mathbf{X}'_+\mathbf{X}_+)^{-1}[\mathbf{X}'_1(\mathbf{I}_{n_1} - \mathbf{H}_1) \quad \mathbf{0}] \\
&= \mathbf{A}_{e_1} - \mathbf{X}_+(\mathbf{X}'_+\mathbf{X}_+)^{-1}[\mathbf{0} \quad \mathbf{0}] \\
&= \mathbf{A}_{e_1}
\end{aligned}$$

vii. Proof similar to part v, above.

viii. Note that $(1 - k_1k_+^{-1}) = k_2k_+^{-1}$,

$$\begin{aligned}
\mathbf{A}_{e_+}\mathbf{A}_{e_p} &= (\mathbf{I}_{n_+} - \mathbf{H}_+)\mathbf{A}_{e_p} \\
&= \mathbf{A}_{e_p} - \mathbf{H}_+\mathbf{A}_{e_p} \\
&= \mathbf{A}_{e_p} - \mathbf{H}_+(\mathbf{A}_{h_1} + \mathbf{A}_{h_2} - \mathbf{A}_{h_+}) \\
&= \mathbf{A}_{e_p} - \mathbf{X}_+(\mathbf{X}'_+\mathbf{X}_+)^{-1}[\mathbf{X}'_1 \quad \mathbf{X}'_2] \begin{bmatrix} (k_1^{-1} - k_+^{-1})\mathbf{X}_1\mathbf{B}\mathbf{X}'_1 & -k_+^{-1}\mathbf{X}_1\mathbf{B}\mathbf{X}'_2 \\ -k_+^{-1}\mathbf{X}_2\mathbf{B}\mathbf{X}'_1 & (k_2^{-1} - k_+^{-1})\mathbf{X}_2\mathbf{B}\mathbf{X}'_2 \end{bmatrix} \\
&= \mathbf{A}_{e_p} - \left\{ \mathbf{X}_+(\mathbf{X}'_+\mathbf{X}_+)^{-1} \times \right. \\
&\quad \left. [(1 - k_1k_+^{-1} - k_2k_+^{-1})\mathbf{X}'_0\mathbf{X}_0\mathbf{B}\mathbf{X}'_1 \quad (1 - k_2k_+^{-1} - k_1k_+^{-1})\mathbf{X}'_0\mathbf{X}_0\mathbf{B}\mathbf{X}'_2] \right\} \\
&= \mathbf{A}_{e_p} - \mathbf{X}_+(\mathbf{X}'_+\mathbf{X}_+)^{-1}[\mathbf{0} \cdot \mathbf{X}'_0\mathbf{X}_0\mathbf{B}\mathbf{X}'_1 \quad \mathbf{0} \cdot \mathbf{X}'_0\mathbf{X}_0\mathbf{B}\mathbf{X}'_2] \\
&= \mathbf{A}_{e_p} - \mathbf{X}_+(\mathbf{X}'_+\mathbf{X}_+)^{-1}[\mathbf{0} \quad \mathbf{0}] \\
&= \mathbf{A}_{e_p}
\end{aligned}$$

d. Lemma 2.1 gives $\mathbf{A}_{h+}\mathbf{A}_{e+} = \mathbf{A}_{h+}\mathbf{A}_{e1} = \mathbf{0}$. Similarly, $\mathbf{A}_{h+}\mathbf{A}_{e2} = \mathbf{0}$.

$\mathbf{A}_{h+}\mathbf{A}_{ep} = \mathbf{0}$:

$$\begin{aligned}\mathbf{A}_{h+}\mathbf{A}_{ep} &= \mathbf{A}_{h+}(\mathbf{A}_{h1} + \mathbf{A}_{h2} - \mathbf{A}_{h+}) \\ &= \mathbf{A}_{h+}\mathbf{A}_{h1} + \mathbf{A}_{h+}\mathbf{A}_{h2} - \mathbf{A}_{h+} \\ &= \mathbf{A}_{h+} - \mathbf{A}_{h+} = \mathbf{0},\end{aligned}$$

$\mathbf{A}_{h+}\mathbf{A}_{eb} = \mathbf{0}$:

$$\begin{aligned}\mathbf{A}_{h+}\mathbf{A}_{eb} &= \mathbf{A}_{h+}(\mathbf{A}_{e+} - \mathbf{A}_{e1} - \mathbf{A}_{e2} - \mathbf{A}_{ep}) \\ &= \mathbf{0} - \mathbf{0} - \mathbf{0} - \mathbf{0} = \mathbf{0},\end{aligned}$$

$\mathbf{A}_{h1}\mathbf{A}_{h2} = \mathbf{0}$:

$$\begin{aligned}\mathbf{A}_{h1}\mathbf{A}_{h2} &= k_1^{-1}k_2^{-1} \begin{bmatrix} \mathbf{X}_1\mathbf{B}\mathbf{X}'_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2\mathbf{B}\mathbf{X}'_2 \end{bmatrix} \\ &= k_1^{-1}k_2^{-1} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \mathbf{0},\end{aligned}$$

$\mathbf{A}_{h1}\mathbf{A}_{e1} = \mathbf{0}$:

$$\begin{aligned}\mathbf{A}_{h1}\mathbf{A}_{e1} &= k_1^{-1} \begin{bmatrix} \mathbf{X}_1\mathbf{B}\mathbf{X}'_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{I}_{n_1} - \mathbf{H}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0}_{n_2 \times n_2} \end{bmatrix} \\ &= k_1^{-1} \begin{bmatrix} \mathbf{X}_1\mathbf{B}\mathbf{X}'_1(\mathbf{I}_{n_1} - \mathbf{H}_1) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \\ &= k_1^{-1} \begin{bmatrix} \mathbf{X}_1\mathbf{B}\mathbf{X}'_1 - \mathbf{X}_1\mathbf{B}\mathbf{X}'_1\mathbf{H}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \\ &= k_1^{-1} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \mathbf{0},\end{aligned}$$

$\mathbf{A}_{h1}\mathbf{A}_{e2} = \mathbf{0}$:

$$\begin{aligned}\mathbf{A}_{h1}\mathbf{A}_{e2} &= k_1^{-1} \begin{bmatrix} \mathbf{X}_1\mathbf{B}\mathbf{X}'_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{0}_{n_1 \times n_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{n_2} - \mathbf{H}_2 \end{bmatrix} \\ &= k_1^{-1} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \mathbf{0},\end{aligned}$$

$\mathbf{A}_{h1}\mathbf{A}_{eb} = \mathbf{0}$ (\mathbf{A}_{h1} idempotent from Section 2.2.3):

$$\begin{aligned}
\mathbf{A}_{h1}\mathbf{A}_{eb} &= \mathbf{A}_{h1}(\mathbf{A}_{e+} - \mathbf{A}_{e1} - \mathbf{A}_{e2} - \mathbf{A}_{ep}) \\
&= \mathbf{A}_{h1}\mathbf{A}_{e+} - \mathbf{A}_{h1}\mathbf{A}_{ep} \\
&= \mathbf{A}_{h1}(\mathbf{I} - \mathbf{H}_+) - \mathbf{A}_{h1}(\mathbf{A}_{h1} + \mathbf{A}_{h2} - \mathbf{A}_{h+}) \\
&= \mathbf{A}_{h1} - \mathbf{A}_{h1}\mathbf{H}_+ - \mathbf{A}_{h1} + \mathbf{A}_{h1}\mathbf{A}_{h+} \\
&= \mathbf{A}_{h1}\mathbf{A}_{h+} - \mathbf{A}_{h1}\mathbf{H}_+ \\
&= k_+^{-1} \begin{bmatrix} \mathbf{X}_1\mathbf{B}\mathbf{X}'_1 & \mathbf{X}_1\mathbf{B}\mathbf{X}'_2 \\ \mathbf{0} & \mathbf{0} \end{bmatrix} - k_+^{-1} \begin{bmatrix} \mathbf{X}_1\mathbf{B}\mathbf{X}'_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} (\mathbf{X}'_+\mathbf{X}_+)^{-1} \mathbf{X}'_+ \\
&= k_+^{-1} \begin{bmatrix} \mathbf{X}_1\mathbf{B}\mathbf{X}'_1 & \mathbf{X}_1\mathbf{B}\mathbf{X}'_2 \\ \mathbf{0} & \mathbf{0} \end{bmatrix} - k_+^{-1} \begin{bmatrix} \mathbf{X}_1\mathbf{B}\mathbf{X}'_0\mathbf{X}_0 \\ \mathbf{0} \end{bmatrix} (\mathbf{X}'_0\mathbf{X}_0)^{-1} [\mathbf{X}'_1 \quad \mathbf{X}'_2] \\
&= k_+^{-1} \begin{bmatrix} \mathbf{X}_1\mathbf{B}\mathbf{X}'_1 & \mathbf{X}_1\mathbf{B}\mathbf{X}'_2 \\ \mathbf{0} & \mathbf{0} \end{bmatrix} - k_+^{-1} \begin{bmatrix} \mathbf{X}_1\mathbf{B}\mathbf{X}'_1 & \mathbf{X}_1\mathbf{B}\mathbf{X}'_2 \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \mathbf{0},
\end{aligned}$$

$$\mathbf{A}_{h2}\mathbf{A}_{e1} = \mathbf{0}:$$

$$\mathbf{A}_{h2}\mathbf{A}_{e1} = k_2^{-1} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2\mathbf{B}\mathbf{X}'_2 \end{bmatrix} \begin{bmatrix} \mathbf{I}_{n_1} - \mathbf{H}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0}_{n_2 \times n_2} \end{bmatrix} = k_2^{-1} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \mathbf{0},$$

$$\mathbf{A}_{h2}\mathbf{A}_{e2} = \mathbf{0}:$$

$$\begin{aligned}
\mathbf{A}_{h2}\mathbf{A}_{e2} &= k_2^{-1} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2\mathbf{B}\mathbf{X}'_2 \end{bmatrix} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{n_2} - \mathbf{H}_2 \end{bmatrix} \\
&= k_2^{-1} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2\mathbf{B}\mathbf{X}'_2(\mathbf{I}_{n_2} - \mathbf{H}_2) \end{bmatrix} \\
&= k_2^{-1} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2\mathbf{B}\mathbf{X}'_2 - \mathbf{X}_2\mathbf{B}\mathbf{X}'_2\mathbf{H}_2 \end{bmatrix} = k_2^{-1} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \mathbf{0},
\end{aligned}$$

$$\mathbf{A}_{h2}\mathbf{A}_{eb} = \mathbf{0} \text{ (idempotency of } \mathbf{A}_{h2} \text{ is established in part f):}$$

$$\begin{aligned}
\mathbf{A}_{h2}\mathbf{A}_{eb} &= \mathbf{A}_{h2}(\mathbf{A}_{e+} - \mathbf{A}_{e1} - \mathbf{A}_{e2} - \mathbf{A}_{ep}) \\
&= \mathbf{A}_{h2}\mathbf{A}_{e+} - \mathbf{A}_{h2}\mathbf{A}_{ep} \\
&= \mathbf{A}_{h2}(\mathbf{I}_{n_+} - \mathbf{H}_+) - \mathbf{A}_{h2}(\mathbf{A}_{h1} + \mathbf{A}_{h2} - \mathbf{A}_{h+}) \\
&= \mathbf{A}_{h2} - \mathbf{A}_{h2}\mathbf{H}_+ - \mathbf{A}_{h2} + \mathbf{A}_{h2}\mathbf{A}_{h+} \\
&= \mathbf{A}_{h2}\mathbf{A}_{h+} - \mathbf{A}_{h2}\mathbf{H}_+ \\
&= k_+^{-1} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{X}_2\mathbf{B}\mathbf{X}'_1 & \mathbf{X}_2\mathbf{B}\mathbf{X}'_2 \end{bmatrix} - k_+^{-1} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2\mathbf{B}\mathbf{X}'_2 \end{bmatrix} \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} (\mathbf{X}'_+\mathbf{X}_+)^{-1} \mathbf{X}'_+ \\
&= k_+^{-1} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{X}_2\mathbf{B}\mathbf{X}'_1 & \mathbf{X}_2\mathbf{B}\mathbf{X}'_2 \end{bmatrix} - k_+^{-1} \begin{bmatrix} \mathbf{0} \\ \mathbf{X}_2\mathbf{B}\mathbf{X}'_0\mathbf{X}_0 \end{bmatrix} (\mathbf{X}'_0\mathbf{X}_0)^{-1} [\mathbf{X}'_1 \quad \mathbf{X}'_2] \\
&= k_+^{-1} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{X}_2\mathbf{B}\mathbf{X}'_1 & \mathbf{X}_2\mathbf{B}\mathbf{X}'_2 \end{bmatrix} - k_+^{-1} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{X}_2\mathbf{B}\mathbf{X}'_1 & \mathbf{X}_2\mathbf{B}\mathbf{X}'_2 \end{bmatrix} = \mathbf{0},
\end{aligned}$$

$$\mathbf{A}_{e1}\mathbf{A}_{e2} = \mathbf{0}:$$

$$\begin{aligned}\mathbf{A}_{e1}\mathbf{A}_{e2} &= \begin{bmatrix} \mathbf{I}_{n_1} - \mathbf{H}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{n_2} - \mathbf{H}_2 \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \mathbf{0},\end{aligned}$$

$$\mathbf{A}_{e1}\mathbf{A}_{ep} = \mathbf{0}:$$

$$\begin{aligned}\mathbf{A}_{e1}\mathbf{A}_{ep} &= \mathbf{A}_{e1}(\mathbf{A}_{h1} + \mathbf{A}_{h2} - \mathbf{A}_{h+}) \\ &= \mathbf{0} + \mathbf{0} - \mathbf{0} = \mathbf{0},\end{aligned}$$

$$\mathbf{A}_{e1}\mathbf{A}_{eb} = \mathbf{0} \text{ (}\mathbf{A}_{e1} \text{ idempotent from Section 2.2.3):}$$

$$\begin{aligned}\mathbf{A}_{e1}\mathbf{A}_{eb} &= \mathbf{A}_{e1}(\mathbf{A}_{e+} - \mathbf{A}_{e1} - \mathbf{A}_{e2} - \mathbf{A}_{ep}) \\ &= \mathbf{A}_{e1}\mathbf{A}_{e+} - \mathbf{A}_{e1} \\ &= \mathbf{A}_{e1} - \mathbf{A}_{e1} = \mathbf{0},\end{aligned}$$

$$\mathbf{A}_{e2}\mathbf{A}_{ep} = \mathbf{0}:$$

$$\begin{aligned}\mathbf{A}_{e2}\mathbf{A}_{ep} &= \mathbf{A}_{e2}(\mathbf{A}_{h1} + \mathbf{A}_{h2} - \mathbf{A}_{h+}) \\ &= \mathbf{0} + \mathbf{0} - \mathbf{0} = \mathbf{0},\end{aligned}$$

$$\mathbf{A}_{e2}\mathbf{A}_{eb} = \mathbf{0} \text{ (idempotency of } \mathbf{A}_{e2} \text{ is established in part f):}$$

$$\begin{aligned}\mathbf{A}_{e2}\mathbf{A}_{eb} &= \mathbf{A}_{e2}(\mathbf{A}_{e+} - \mathbf{A}_{e1} - \mathbf{A}_{e2} - \mathbf{A}_{ep}) \\ &= \mathbf{A}_{e2}\mathbf{A}_{e+} - \mathbf{A}_{e2} \\ &= \mathbf{A}_{e2} - \mathbf{A}_{e2} = \mathbf{0},\end{aligned}$$

$$\mathbf{A}_{ep}\mathbf{A}_{eb} = \mathbf{0}:$$

$$\begin{aligned}\mathbf{A}_{ep}\mathbf{A}_{eb} &= (\mathbf{A}_{h1} + \mathbf{A}_{h2} - \mathbf{A}_{h+})(\mathbf{A}_{e+} - \mathbf{A}_{e1} - \mathbf{A}_{e2} - \mathbf{A}_{ep}) \\ &= \mathbf{A}_{h1}\mathbf{A}_{e+} - \mathbf{A}_{h1}\mathbf{A}_{ep} + \mathbf{A}_{h2}\mathbf{A}_{e+} - \mathbf{A}_{h2}\mathbf{A}_{ep} \\ &= (\mathbf{A}_{h1}\mathbf{A}_{e+} - \mathbf{A}_{h1}\mathbf{A}_{ep}) + (\mathbf{A}_{h2}\mathbf{A}_{e+} - \mathbf{A}_{h2}\mathbf{A}_{ep}) \\ &= \mathbf{A}_{h1}\mathbf{A}_{eb} + \mathbf{A}_{h2}\mathbf{A}_{eb} \\ &= \mathbf{0} + \mathbf{0} = \mathbf{0}.\end{aligned}$$

e. Idempotency of \mathbf{A}_{h+} , \mathbf{A}_{h1} , \mathbf{A}_{e+} , and \mathbf{A}_{e1} are given in Lemma 2.1. Also, since labeling order of data is arbitrary, proofs about middle matrices for first sample apply to second sample without loss of generality, so \mathbf{A}_{h2} and \mathbf{A}_{e2} are also idempotent. This leaves showing idempotency of \mathbf{A}_{eb} and \mathbf{A}_{ep} . The results below use properties derived earlier in this lemma.

$$\begin{aligned}
\mathbf{A}_{ep}\mathbf{A}_{ep} &= (\mathbf{A}_{h1} + \mathbf{A}_{h2} - \mathbf{A}_{h+})(\mathbf{A}_{h1} + \mathbf{A}_{h2} - \mathbf{A}_{h+}) \\
&= \mathbf{A}_{h1} - \mathbf{A}_{h1}\mathbf{A}_{h+} + \mathbf{A}_{h2} - \mathbf{A}_{h2}\mathbf{A}_{h+} - \mathbf{A}_{h+}\mathbf{A}_{h1} - \mathbf{A}_{h+}\mathbf{A}_{h2} + \mathbf{A}_{h+} \\
&= \mathbf{A}_{h1} + \mathbf{A}_{h2} + \mathbf{A}_{h+} - (\mathbf{A}_{h1}\mathbf{A}_{h+} + \mathbf{A}_{h+}\mathbf{A}_{h1}) - (\mathbf{A}_{h2}\mathbf{A}_{h+} + \mathbf{A}_{h+}\mathbf{A}_{h2}) \\
&= \mathbf{A}_{h1} + \mathbf{A}_{h2} - \mathbf{A}_{h+} - (k_+^{-1}k_1\mathbf{A}_{h1} - k_+^{-1}k_2\mathbf{A}_{h2}) + (k_+^{-1}k_1\mathbf{A}_{h1} - k_+^{-1}k_2\mathbf{A}_{h2}) \\
&= \mathbf{A}_{h1} + \mathbf{A}_{h2} - \mathbf{A}_{h+} \\
&= \mathbf{A}_{ep}
\end{aligned}$$

$$\begin{aligned}
\mathbf{A}_{eb}\mathbf{A}_{eb} &= (\mathbf{A}_{e+} - \mathbf{A}_{e1} - \mathbf{A}_{e2} - \mathbf{A}_{ep})(\mathbf{A}_{e+} - \mathbf{A}_{e1} - \mathbf{A}_{e2} - \mathbf{A}_{ep}) \\
&= \mathbf{A}_{e+} - \mathbf{A}_{e+}\mathbf{A}_{e1} - \mathbf{A}_{e+}\mathbf{A}_{e2} - \mathbf{A}_{e+}\mathbf{A}_{ep} - \mathbf{A}_{e1}\mathbf{A}_{e+} + \\
&\quad \mathbf{A}_{e1} - \mathbf{A}_{e2}\mathbf{A}_{e+} + \mathbf{A}_{e2} - \mathbf{A}_{ep}\mathbf{A}_{e+} + \mathbf{A}_{ep} \\
&= \mathbf{A}_{e+} - \mathbf{A}_{e1} - \mathbf{A}_{e2} - \mathbf{A}_{ep} - \mathbf{A}_{e1} + \mathbf{A}_{e1} - \mathbf{A}_{e2} + \mathbf{A}_{e2} - \mathbf{A}_{ep} + \mathbf{A}_{ep} \\
&= \mathbf{A}_{e+} - \mathbf{A}_{e1} - \mathbf{A}_{e2} - \mathbf{A}_{ep} \\
&= \mathbf{A}_{eb}
\end{aligned}$$

f. Ranks of \mathbf{A}_{h+} , \mathbf{A}_{h1} , \mathbf{A}_{e+} , and \mathbf{A}_{e1} are given in Lemma 2.1. Also, since the labeling order of data is arbitrary, it follows that \mathbf{A}_{h2} and \mathbf{A}_{e2} have ranks a and $n_2 - r$, respectively. This leaves showing rank for \mathbf{A}_{eb} and \mathbf{A}_{ep} .

$$\begin{aligned}
\mathbf{A}_{ep} &= \mathbf{A}_{h1} + \mathbf{A}_{h2} - \mathbf{A}_{h+} \\
&= \begin{bmatrix} (k_1^{-1} - k_+^{-1})\mathbf{X}_1\mathbf{B}\mathbf{X}'_1 & -k_+^{-1}\mathbf{X}_1\mathbf{B}\mathbf{X}'_2 \\ -k_+^{-1}\mathbf{X}_2\mathbf{B}\mathbf{X}'_1 & (k_2^{-1} - k_+^{-1})\mathbf{X}_2\mathbf{B}\mathbf{X}'_2 \end{bmatrix} \\
&= \begin{bmatrix} \left(\frac{k_+ - k_1}{k_1 k_+}\right)(\mathbf{1}_{k_1} \otimes \mathbf{X}_0)\mathbf{B}(\mathbf{1}'_{k_1} \otimes \mathbf{X}'_0) & -k_+^{-1}(\mathbf{1}_{k_1} \otimes \mathbf{X}_0)\mathbf{B}(\mathbf{1}'_{k_2} \otimes \mathbf{X}'_0) \\ -k_+^{-1}(\mathbf{1}_{k_2} \otimes \mathbf{X}_0)\mathbf{B}(\mathbf{1}'_{k_1} \otimes \mathbf{X}'_0) & \left(\frac{k_+ - k_2}{k_2 k_+}\right)(\mathbf{1}_{k_2} \otimes \mathbf{X}_0)\mathbf{B}(\mathbf{1}'_{k_2} \otimes \mathbf{X}'_0) \end{bmatrix} \\
&= k_+^{-1} \begin{bmatrix} (k_2/k_1)(\mathbf{1}_{k_1} \otimes \mathbf{X}_0\mathbf{B})(\mathbf{1}'_{k_1} \otimes \mathbf{X}'_0) & -(\mathbf{1}_{k_1} \otimes \mathbf{X}_0\mathbf{B})(\mathbf{1}'_{k_2} \otimes \mathbf{X}'_0) \\ -(\mathbf{1}_{k_2} \otimes \mathbf{X}_0\mathbf{B})(\mathbf{1}'_{k_1} \otimes \mathbf{X}'_0) & (k_1/k_2)(\mathbf{1}_{k_2} \otimes \mathbf{X}_0\mathbf{B})(\mathbf{1}'_{k_2} \otimes \mathbf{X}'_0) \end{bmatrix} \\
&= k_+^{-1} \begin{bmatrix} (k_2/k_1)(\mathbf{1}_{k_1}\mathbf{1}'_{k_1} \otimes \mathbf{X}_0\mathbf{B}\mathbf{X}'_0) & -(\mathbf{1}_{k_1}\mathbf{1}'_{k_2} \otimes \mathbf{X}_0\mathbf{B}\mathbf{X}'_0) \\ -(\mathbf{1}_{k_2}\mathbf{1}'_{k_1} \otimes \mathbf{X}_0\mathbf{B}\mathbf{X}'_0) & (k_1/k_2)(\mathbf{1}_{k_2}\mathbf{1}'_{k_2} \otimes \mathbf{X}_0\mathbf{B}\mathbf{X}'_0) \end{bmatrix} \\
&= k_+^{-1} \begin{bmatrix} (k_2/k_1)(\mathbf{1}_{k_1}\mathbf{1}'_{k_1}) & -(\mathbf{1}_{k_1}\mathbf{1}'_{k_2}) \\ -(\mathbf{1}_{k_2}\mathbf{1}'_{k_1}) & (k_1/k_2)(\mathbf{1}_{k_2}\mathbf{1}'_{k_2}) \end{bmatrix} \otimes \mathbf{X}_0\mathbf{B}\mathbf{X}'_0
\end{aligned}$$

Multiplying the first k_1 rows of the left matrix by $-(k_1/k_2)$ creates a matrix of identical rows, and so the left matrix has rank 1. Now, $\mathbf{A}_{h+} = \mathbf{1}_{k_+}\mathbf{1}'_{k_+} \otimes \mathbf{X}_0\mathbf{B}\mathbf{X}'_0$ is of rank a and

$\mathbf{1}_{k_+} \mathbf{1}'_{k_+}$ is clearly rank 1. The fact that the rank of a Kronecker product is the product of the ranks, allows showing that $\mathbf{X}_0 \mathbf{B} \mathbf{X}'_0$ is rank a which implies that \mathbf{A}_{ep} is also rank a .

The matrix \mathbf{A}_{eb} is defined as $\mathbf{A}_{e+} - \mathbf{A}_{e1} - \mathbf{A}_{e2} - \mathbf{A}_{ep}$. Therefore $\mathbf{A}_{e+} = \mathbf{A}_{e1} + \mathbf{A}_{e2} + \mathbf{A}_{ep} + \mathbf{A}_{eb}$ decomposes \mathbf{A}_{e+} into four symmetric, idempotent pieces. Now since \mathbf{A}_{e+} is also idempotent and symmetric, a matrix decomposition theorem from Muller and Stewart (2006; Theorem 9.16, page 187) to show that the sum of the ranks of the pieces is equal to the rank of \mathbf{A}_{e+} . This then implies for $r = \text{rank}(\mathbf{X}_0)$ that

$$\begin{aligned} \text{rank}(\mathbf{A}_{eb}) &= \text{rank}(\mathbf{A}_{e+}) - \text{rank}(\mathbf{A}_{e1}) - \text{rank}(\mathbf{A}_{e2}) - \text{rank}(\mathbf{A}_{ep}) \\ &= (n_+ - r) - (n_1 - r) - (n_2 - r) - a \\ &= r - a \end{aligned}$$

□

Lemma 4.3 Proof. For $a > 1$ and $X_1^2 \sim \chi^2(a, \lambda_+)$ defined as in equation 4.41 and using results from equations 4.34 and 4.46-4.49,

$$\begin{aligned} \widehat{\delta}_+ &= \mathbf{y}'_+ \mathbf{A}_{h+} \mathbf{y}_+ \\ &= \mathbf{y}'_{h+} \mathbf{y}_{h+} \\ &= \mathbf{1}'_2 \mathbf{S}_1 \mathbf{1}_2 \\ &= S_{11} + S_{12} + S_{21} + S_{22} \\ &= \sigma^2 X_1^2 \end{aligned}$$

□

Lemma 4.4 Proof. For $a > 1$ and X_1^2 , X_2^2 , and Z defined as in equations 4.41-4.43, using results from equations 4.33 and 4.46,

$$\begin{aligned} \widehat{\delta}_1 &= \mathbf{y}'_+ \mathbf{A}_{h1} \mathbf{y}_+ \\ &= \mathbf{y}'_{h1} \mathbf{y}_{h1} \\ &= (n_+/n_1) S_{11} \\ &= \sigma^2 n_+^{-1} \left[(\sqrt{n_1} X_1 + \sqrt{n_2} Z)^2 + n_2 X_2^2 \right] \end{aligned}$$

□

Lemma 4.5 Proof. For $a > 1$ and X_1^2, X_2^2 , and Z defined as in equations 4.41-4.43, using results from equations 4.33 and 4.49,

$$\begin{aligned}
\widehat{\delta}_2 &= \mathbf{y}'_+ \mathbf{A}_{h2} \mathbf{y}_+ \\
&= \mathbf{y}'_{h2} \mathbf{y}_{h2} \\
&= (n_+/n_2) S_{11} \\
&= \sigma^2 n_+^{-1} \left[(\sqrt{n_1} X_1 - \sqrt{n_2} Z)^2 + n_1 X_2^2 \right] \quad \square
\end{aligned}$$

The following two lemmas will be useful in the proofs for Lemma 4.6 and Theorem 4.2 which follow. They are distributional results involving a standard Gaussian and it's square.

Lemma B.1 For $\{h, l\} \geq 0$ and $Z \sim \mathcal{N}(0, 1)$, the following holds.

$$\begin{aligned}
&\Pr\{(l + Z)^2 \leq h^2\} \\
&= \Pr\{|l + Z| \leq h\} \\
&= \Pr\{-h \leq l + Z \leq h\} \\
&= \Pr\{-(h + l) \leq Z \leq (h - l)\} \\
&= \Phi[(h - l)] - \Phi[-(h + l)] \quad \square
\end{aligned}$$

Lemma B.2 For $\{d, h\} \geq 0$ and $Z \sim \mathcal{N}(0, 1)$, the following holds

$$\begin{aligned}
&\Pr\{(-d \leq Z \leq b) \cap (Z^2 \geq h^2)\} \\
&= \Pr\{(-d \leq Z \leq b) \cap (|Z| \geq h)\} \\
&= \Pr\{(-d \leq Z \leq b) \cap [(Z \leq -h) \cup (Z \geq h)]\} \\
&= \Pr\{(-d \leq Z \leq b) \cap (Z \leq -h)\} + \Pr\{(-d \leq Z \leq b) \cap (Z \geq h)\} \\
&= \Pr\{-d \leq Z \leq \min(b, -h)\} + \Pr\{h \leq Z \leq b\} \\
&= \max\{\Phi[\min(b, -h)] - \Phi(-d), 0\} + \max\{\Phi(b) - \Phi(h), 0\} \quad \square
\end{aligned}$$

Lemma 4.6 Proof. Let $u_1 = c_+ g_+$, $u_2 = c_+(g_1/n_2 + g_+)$, $u_3(p_1) = c_+^{-1} p_1 - g_+$, $u_4 = g_1/n_2$, $b(p_1, p_2) = \sqrt{g_1/n_2 - p_2} - \sqrt{p_1 n_1/n_2}$, $f_{p_1} = f_{X_1^2}(p_1)$, $f_{p_2} = f_{X_2^2}(p_2)$, $d(p_1, p_2) = \sqrt{g_1/n_2 - p_2} + \sqrt{p_1 n_1/n_2}$, $h(p_1, p_2) = \sqrt{c_+^{-1} p_1 - p_2 - g_+}$. For strictly positive $\{g_1, g_+, c_+, n_1, n_2\}$, integer $a > 1$, $Z \sim \mathcal{N}(0, 1)$, $X_1^2 \sim \chi^2(a, \lambda_+)$, $X_2^2 \sim \chi^2(a - 1)$ with Z, X_1^2, X_2^2 mutually independent, the following holds.

$$\begin{aligned}
& \Pr \left\{ \left[(\sqrt{n_1}X_1 + \sqrt{n_2}Z)^2 + n_2X_2^2 \leq g_1 \right] \cap [c_+^{-1}X_1^2 - X_2^2 - Z^2 \leq g_+] \right\} \\
&= \int_0^\infty f_{p_1} \Pr \left\{ \left[(\sqrt{p_1n_1/n_2} + Z)^2 + X_2^2 \leq g_1/n_2 \right] \cap (c_+^{-1}p_1 - X_2^2 - g_+ \leq Z^2) \right\} dp_1 \\
&= \int_0^\infty \int_0^{g_1/n_2} f_{p_1} f_{p_2} \\
&\quad \times \Pr \left\{ \left[(\sqrt{p_1n_1/n_2} + Z)^2 + p_2 \leq g_1/n_2 \right] \cap (c_+^{-1}p_1 - p_2 - g_+ \leq Z^2) \right\} dp_2 dp_1 \\
&= \int_0^{c_+g_+} \int_0^{g_1/n_2} f_{p_1} f_{p_2} \Pr \left\{ (\sqrt{p_1n_1/n_2} + Z)^2 \leq g_1/n_2 - p_2 \right\} dp_2 dp_1 \\
&+ \int_{c_+g_+}^{c_+(g_1/n_2+g_+)} \int_0^{c_+^{-1}p_1-g_+} f_{p_1} f_{p_2} \\
&\quad \times \Pr \left\{ \left[(\sqrt{p_1n_1/n_2} + Z)^2 \leq g_1/n_2 - p_2 \right] \cap (c_+^{-1}p_1 - p_2 - g_+ \leq Z^2) \right\} dp_2 dp_1 \\
&+ \int_{c_+g_+}^{c_+(g_1/n_2+g_+)} \int_{c_+^{-1}p_1-g_+}^{g_1/n_2} f_{p_1} f_{p_2} \Pr \left\{ (\sqrt{p_1n_1/n_2} + Z)^2 \leq g_1/n_2 - p_2 \right\} dp_2 dp_1 \\
&+ \int_{c_+(g_1/n_2+g_+)}^\infty \int_0^{g_1/n_2} f_{p_1} f_{p_2} \\
&\quad \times \Pr \left\{ \left[(\sqrt{p_1n_1/n_2} + Z)^2 \leq g_1/n_2 - p_2 \right] \cap (c_+^{-1}p_1 - p_2 - g_+ \leq Z^2) \right\} dp_2 dp_1
\end{aligned}$$

$$\begin{aligned}
&= \int_0^{u_1} \int_0^{u_4} f_{p_1} f_{p_2} \Pr\{-d(p_1, p_2) \leq Z \leq b(p_1, p_2)\} dp_2 dp_1 \\
&+ \int_{u_1}^{u_2} \int_0^{u_3(p_1)} f_{p_1} f_{p_2} \Pr\{[-d(p_1, p_2) \leq Z \leq b(p_1, p_2)] \cap (Z^2 \geq h^2(p_1, p_2))\} dp_2 dp_1 \\
&+ \int_{u_1}^{u_2} \int_{u_3(p_1)}^{u_4} f_{p_1} f_{p_2} \Pr\{-d(p_1, p_2) \leq Z \leq b(p_1, p_2)\} dp_2 dp_1 \\
&+ \int_{u_2}^{\infty} \int_0^{u_4} f_{p_1} f_{p_2} \Pr\{[-d(p_1, p_2) \leq Z \leq b(p_1, p_2)] \cap (Z^2 \geq h^2(p_1, p_2))\} dp_2 dp_1 \\
&= \int_0^{u_1} \int_0^{u_4} f_{p_1} f_{p_2} \{\Phi[b(p_1, p_2)] - \Phi[-d(p_1, p_2)]\} dp_2 dp_1 \\
&+ \int_{u_1}^{u_2} \int_0^{u_3(p_1)} f_{p_1} f_{p_2} \left[\max\left(\Phi\{\min[-h(p_1, p_2), b(p_1, p_2)]\} - \Phi[-d(p_1, p_2)], 0\right) \right. \\
&\quad \left. + \max\{\Phi[b(p_1, p_2)] - \Phi[h(p_1, p_2)], 0\} \right] dp_2 dp_1 \\
&+ \int_{u_1}^{u_2} \int_{u_3(p_1)}^{u_4} f_{p_1} f_{p_2} \{\Phi[b(p_1, p_2)] - \Phi[-d(p_1, p_2)]\} dp_2 dp_1 \\
&+ \int_{u_2}^{\infty} \int_0^{u_4} f_{p_1} f_{p_2} \left[\max\left(\Phi\{\min[-h(p_1, p_2), b(p_1, p_2)]\} - \Phi[-d(p_1, p_2)], 0\right) \right. \\
&\quad \left. + \max\{\Phi[b(p_1, p_2)] - \Phi[h(p_1, p_2)], 0\} \right] dp_2 dp_1 \quad \square
\end{aligned}$$

Theorem 4.1 Proof. $a > 1$, $Z \sim \mathcal{N}(0, 1)$, $X_1^2 \sim \chi^2(a, \lambda_+)$, $X_2^2 \sim \chi^2(a - 1)$ with Z , X_1^2 , X_2^2 mutually independent. $E_+ = \nu_+ \widehat{\sigma}_+^2 / \sigma^2 = \mathbf{y}_+ \mathbf{A}_{e1} \mathbf{y}'_+ / \sigma^2$,
 $E_1 = \nu_1 \widehat{\sigma}_1^2 / \sigma^2 = \mathbf{y}_+ \mathbf{A}_{e1} \mathbf{y}'_+ / \sigma^2$, $E_2 = \mathbf{y}_+ (\mathbf{A}_{e2} + \mathbf{A}_{eb}) \mathbf{y}'_+ / \sigma^2$, $E_3 = \mathbf{y}_+ \mathbf{A}_{ep} \mathbf{y}'_+ / \sigma^2$.
 $G_s = \widehat{\delta}_s / \sigma^2 = \mathbf{y}_+ \mathbf{A}_{hs} \mathbf{y}'_+ / \sigma^2$ and $c_s = a f_s / \nu_s$ for $s \in \{1, 2, +\}$.
 $f_{E_1}(t_1) = f_{\chi^2}(t_1; \nu_1) / [F_{\chi^2}(q_2; \nu_1) - F_{\chi^2}(q_1; \nu_1)]$, $f_{E_2}(t_2) = f_{\chi^2}(t_2; n_2 - a)$

$$\begin{aligned}
& F_{F_1, E_+ | N_+}(f_1, f_+) \\
&= \Pr_{N_+} \left\{ \left[\left(\widehat{\delta}_1 / a \right) / \widehat{\sigma}_1^2 \leq f_1 \right] \cap \left[\left(\widehat{\delta}_+ / a \right) / \widehat{\sigma}_+^2 \leq f_+ \right] \right\} \\
&= \Pr_{N_+} \{ (G_1 \leq c_1 E_1) \cap (G_+ \leq c_+ E_+) \} \\
&= \Pr_{N_+} \{ (G_1 \leq c_1 E_1) \cap [c_+^{-1} G_+ \leq E_1 + E_2 + E_3] \} \\
&= \Pr_{N_+} \{ (G_1 \leq c_1 E_1) \cap [c_+^{-1} G_+ - E_3 \leq E_1 + E_2] \} \\
&= \int_{q_1}^{q_2} f_{E_1}(t_1) \Pr_{N_+} \{ (G_1 \leq c_1 t_1) \cap [c_+^{-1} G_+ - E_3 \leq t_1 + E_2] \} dt_1 \\
&= \int_{q_1}^{q_2} \int_0^\infty f_{E_1}(t_1) f_{E_2}(t_2) \Pr_{N_+} \{ (G_1 \leq c_1 t_1) \cap [c_+^{-1} G_+ - E_3 \leq t_1 + t_2] \} dt_2 dt_1 \\
&= \int_{q_1}^{q_2} \int_0^\infty f_{E_1}(t_1) f_{E_2}(t_2) \\
&\quad \times \Pr_{N_+} \{ (G_1 \leq c_1 t_1) \cap [c_+^{-1} G_+ - (G_1 + G_2 - G_+) \leq t_1 + t_2] \} dt_2 dt_1 \\
&= \int_{q_1}^{q_2} \int_0^\infty f_{E_1}(t_1) f_{E_2}(t_2) \\
&\quad \times \Pr_{N_+} \{ (G_1 \leq c_1 t_1) \cap [(1 + c_+^{-1}) G_+ - (G_1 + G_2) \leq t_1 + t_2] \} dt_2 dt_1 \\
&= \int_{q_1}^{q_2} \int_0^\infty f_{E_1}(t_1) f_{E_2}(t_2) \Pr_{N_+} \left\{ \left(n_+^{-1} \left[(\sqrt{n_1} X_1 + \sqrt{n_2} Z)^2 + n_2 X_2^2 \right] \leq c_1 t_1 \right) \right. \\
&\quad \left. \cap \left[(1 + c_+^{-1}) X_1^2 - (X_1^2 + X_2^2 + Z^2) \leq t_1 + t_2 \right] \right\} dt_2 dt_1 \\
&= \int_{q_1}^{q_2} \int_0^\infty f_{E_1}(t_1) f_{E_2}(t_2) \Pr_{N_+} \left\{ \left[(\sqrt{n_1} X_1 + \sqrt{n_2} Z)^2 + n_2 X_2^2 \leq n_+ c_1 t_1 \right] \right. \\
&\quad \left. \cap \left[c_+^{-1} X_1^2 - X_2^2 - Z^2 \leq t_1 + t_2 \right] \right\} dt_2 dt_1 \quad \square
\end{aligned}$$

REFERENCES

- Anscombe, F. J. (1953). Sequential Estimation. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, **15**, 1-29.
- Arghami, N. R., and Billard, L. (1992). Some sequential analogs of steins 2-sample test. *Theory of Probability and its Applications*, **37**, 115-117.
- Armitage, P. (1975). *Sequential Medical Trials*, Oxford: Blackwell.
- Baker A. G. (1950). Properties of some tests in sequential analysis. *Biometrika*, **37**, 334-346.
- Bauer, P., and Köhne, K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics*, **50**(4), 1029-1041.
- Bernardo, M. V. P., and Ibrahim, J. G. (2000). Group sequential designs for cure rate models with early stopping in favour of the null hypothesis. *Statistics in Medicine*, **19**, 3023-3035.
- Birkett, M. A., and Day, S. J. (1994). Internal pilot-studies for estimating sample-size. *Statistics in Medicine*, **13**, 2455-2463.
- Bolland, K., Sooriyarachchi, M. R., and Whitehead, J. (1998). Sample size review in a head injury trial with ordered categorical responses. *Statistics in Medicine*, **17**, 2835-2847.
- Burman, C. F., and Sonesson, C. (2006). Are flexible designs sound? *Biometrics*, **62**, 664-669.
- Choi, S. C., and Pepple, P. A. (1989). Monitoring clinical-trials based on predictive probability of significance. *Biometrics*, **45**, 317-323.
- Choi, S. C., Smith, P. J., and Becker, D. P. (1985). Early decision in clinical-trials when the treatment differences are small - experience of a controlled trial in head trauma. *Controlled Clinical Trials*, **6**, 280-288.
- Coffey, C. S., and Muller, K. E. (2003). Properties of internal pilots with the univariate approach to repeated measures. *Statistics in Medicine*, **22**, 2469-2485.
- Coffey, C. S., and Muller, K. E. (2001). Controlling test size while gaining the benefits of an internal pilot design. *Biometrics*, **57**, 625-631.
- Coffey, C. S., and Muller, K. E. (2000). Some distributions and their implications for an internal pilot study with a univariate linear model. *Communications in Statistics-Theory and Methods*, **29**, 2677-2691.
- Coffey, C. S., and Muller, K. E. (1999). Exact test size and power of a gaussian error linear model for an internal pilot study. *Statistics in Medicine*, **18**, 1199-1214.

- Cui, L., Hung, H. M. J., and Wang, S. J. (1999). Modification of sample size in group sequential clinical trials. *Biometrics*, **55**, 853-857.
- Denne, J. S. (2001). Sample size recalculation using conditional power. *Statistics in Medicine*, **20**, 2645-2660.
- Denne, J. S., and Jennison, C. (2000). A group sequential t-test with updating of sample size. *Biometrika*, **87**, 125-134.
- Denne, J. S., and Jennison, C. (1999). Estimating the sample size for a t-test using an internal pilot. *Statistics in Medicine*, **18**, 1575-1585.
- Dmitrienko, A., and Wang, M. D. (2006). Bayesian predictive approach to interim monitoring in clinical trials. *Statistics in Medicine*, **25**, 2178-2195.
- Dmitrienko, A., Molenberghs, G., Chuang-Stein, C., and Offen, W. (2005). *Analysis of Clinical Trials Using SAS: A Practical Guide*, The SAS Institute, Cary, NC.
- Facey, K. M. (1992). A sequential procedure for a phase-ii efficacy trial in hypercholesterolemia. *Controlled Clinical Trials*, **13**, 122-133.
- Friede, T., and Kieser, M. (2006). Sample size recalculation in internal pilot study designs: A review. *Biometrical Journal*, **48**, 537-555.
- Friede, T., and Kieser, M. (2001). A comparison of methods for adaptive sample size adjustment. *Statistics in Medicine*, **20**, 3861-3873.
- Gallo, P., Chuang-Stein, C., Dragalin, V., Gaydos, B., Krams, M., and Pinheiro, J. (2006). Adaptive designs in clinical drug development - an executive summary of the PhRMA working group. *Journal of Biopharmaceutical Statistics*, **16**, 275-283.
- Geisser, S. (1992). On the curtailment of sampling. *Canadian Journal of Statistics-Revue Canadienne De Statistique*, **20**, 297-309.
- Glueck, D.H. and Muller, K.E. (2001). On the expected values of sequences of functions. *Communications in Statistics-Theory and Methods*, **30**, 363-369
- Gould, A. L. (2001). Sample size re-estimation: Recent developments and practical considerations. *Statistics in Medicine*, **20**, 2625-2643.
- Gould, A. L. (1995). Planning and revising the sample-size for a trial. *Statistics in Medicine*, **14**, 1039-1051.
- Gould, A.L. (1983). Abandoning lost causes (Early termination of unproductive clinical trials). *Proceedings of Biopharmaceutical Section, American Statistical Association*, Washington, D.C., 31-41.

- Gould, A.L., and Shih, W. J. (1998). Modifying the design of ongoing trials without unblinding. *Statistics in Medicine*, **17**, 89-100.
- Gould, A.L., and Shih, W.J. (1992). Sample-size reestimation without unblinding for normally distributed outcomes with unknown-variance. *Communications in Statistics-Theory and Methods*, **21**, 2833-2853.
- Gupta, A.K. and Nagar, D.K. (2000) *Matrix Variate Distributions*, Boca Raton: Chapman & Hall/CRC.
- Gurka M.J., Coffey C.S., and Muller K.E. (2007) Internal pilots for a class of linear mixed models with Gaussian and compound symmetric data, *Statistics in Medicine*, **26**, 4083–4099.
- Hall, W.J. (1962). Some sequential analogs of steins 2-stage test. *Biometrika*, **49**, 367-378.
- Haybittle, J.L. (1971). Repeated assessment of results in clinical trials of cancer treatment. *British Journal of Radiology*, **44**, 793-and.
- Hewett, J. E., and Spurrier, J.D. (1983). A survey of 2 stage tests of hypotheses - theory and application. *Communications in Statistics-Theory and Methods*, **12**, 2307-2425.
- Hochberg, Y., and Marcus, R. (1983). 2-phase tests on a normal-mean when variance is unknown. *Journal of Statistical Planning and Inference*, **7**, 233-242.
- Hwang, I.K., Shih, W.J., and Decani, J. S. (1990). Group sequential designs using a family of type-i error-probability spending functions. *Statistics in Medicine*, **9**, 1439-1445.
- ICH Guideline E9. (1998). *International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use*. ICH Topic E9: Statistical principles for clinical trials. ICH Technical Coordination, EMEA, London.
- Jennison, C., and Turnbull, B.W. (2006). Adaptive and nonadaptive group sequential tests. *Biometrika*, **93**, 1-21.
- Jennison, C., and Turnbull, B.W. (2006). Efficient group sequential designs when there are several effect sizes under consideration. *Statistics in Medicine*, **25**, 917-932.
- Jennison, C., and Turnbull, B.W. (2003). Mid-course sample size modification in clinical trials based on the observed treatment effect. *Statistics in Medicine*, **22**, 971-993.
- Jennison, C., and Turnbull, B.W. (2001). On group sequential tests for data in unequally sized groups and with unknown variance. *Journal of Statistical Planning and Inference*, **96**, 263-288.
- Jennison, C., and Turnbull, B.W. (2000), *Group Sequential Methods with Applications to Clinical Trials*, Boca Raton: Chapman & Hall/CRC.

- Jennison, C., and Turnbull, B.W. (1997). Distribution theory of group sequential t, chi-2, and F-tests for general linear models. *Sequential Analysis*, **16**, 295-317.
- Jennison, C., and Turnbull, B.W. (1991). Exact calculations for sequential T-tests, chi-2-tests and F-tests. *Biometrika*, **78**, 133-141.
- Jennison, C., and Turnbull, B. W. (1989). Interim analyses - the repeated confidence-interval approach. *Journal of the Royal Statistical Society Series B-Methodological*, **51**, 305-361.
- Johnson J.L. (2007). Fixed Effects Inference for Clustered Data in Gaussian Linear Models. Dr.PH. Dissertation, The University of North Carolina at Chapel Hill, Department of Biostatistics.
- Johnson N.L., Kotz S., and Balakrishnan N. (1994). *Continuous Univariate Distributions-1 (2nd ed.)*. New York: Wiley.
- Johnson N.L., Kotz S., and Balakrishnan N. (1995). *Continuous Univariate Distributions-2 (2nd ed.)*. New York: Wiley.
- Kairalla J.A., Coffey C.S., and Muller K.E. (2007). GLUMIP 2.0: SAS/IML software for planning internal pilots. In Revision.
- Kieser, M., and Friede, T. (2003). Simple procedures for blinded sample size adjustment that do not affect the type I error rate. *Statistics in Medicine*, **22**, 3571-3581.
- Kieser, M., and Friede, T. (2000). Re-calculating the sample size in internal pilot study designs with control of the type I error rate. *Statistics in Medicine*, **19**, 901-911.
- Kim, K., and Demets, D. L. (1987). Design and analysis of group sequential-tests based on the type-i error spending rate-function. *Biometrika*, **74**, 149-154.
- Lake, S., Kammann, E., Klar, N., and Betensky, R. (2002). Sample size re-estimation in cluster randomization trials. *Statistics in Medicine*, **21**, 1337-1350.
- Lan, K. K. G., and Demets, D. L. (1983). Discrete sequential boundaries for clinical-trials. *Biometrika*, **70**, 659-663.
- Lan, K. K. G., Simon, R., and Halperin, M. (1982). Stochastically curtailed tests in long-term clinical trials. *Communications in Statistics C*, **1**, 207-219.
- Lehmacher, W., and Wassmer, G. (1999). Adaptive sample size calculations in group sequential trials. *Biometrics*, **55**, 1286-1290.
- Mehta, C.R. and Tsiatis, A.A. (2001) Flexible sample size considerations using information-based interim monitoring. *Drug Information Journal*, **35**: 1095-1112

- Miller, F. (2005). Variance estimation in clinical studies with interim sample size reestimation. *Biometrics*, **61**, 355-361.
- Morgan, C. C. (2003). Sample size re-estimation in group-sequential response-adaptive clinical trials. *Statistics in Medicine*, **22**, 3843-3857.
- Muller K.E., Edwards L.J., Simpson, S.L., and Taylor D.J. (2007) Statistical tests with accurate size and power for balanced linear mixed models, *Statistics in Medicine*, **26**, 3639–3660.
- Muller, K.E., LaVange, L.M., Ramey, S.L., Ramey, C.T. (1992). Power calculations for general linear multivariate models including repeated measures applications. *Journal of the American Statistical Association*, **87**: 1209-1226.
- Muller, K. E., and Stewart, P. (2006). *Linear Model Theory: Univariate, Multivariate, and Mixed Models*, Wiley.
- O'Brien, P. C., and Fleming, T. R. (1979). Multiple testing procedure for clinical-trials. *Biometrics*, **35**, 549-556.
- Pampallona, S., and Tsiatis, A. A. (1994). Group sequential designs for one-sided and 2-sided hypothesis-testing with provision for early stopping in favor of the null hypothesis. *Journal of Statistical Planning and Inference*, **42**, 19-35.
- Park S. (2007) Accounting for Bias and Uncertainty in Power for Multivariate Gaussian Linear Models. Ph.D. Dissertation, University of North Carolina at Chapel Hill, Department of Biostatistics.
- Peto, R., Pike, M. C., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S. V., Mantel, N., McPherson, K., Peto, J., and Smith, P. G. (1976). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and Design. *British Journal of Cancer*, **34**, 585-612.
- Pocock, S. J. (1977). Group sequential methods in design and analysis of clinical-trials. *Biometrika*, **64**, 191-200.
- Proschan, M. A., Lan, G. K. K., and Wittes, J. T. (2006). *Statistical Monitoring of Clinical Trials*, Springer.
- Proschan, M. A. (2005). Two-stage sample size re-estimation based on a nuisance parameter: A review. *Journal of Biopharmaceutical Statistics*, **15**, 559-574.
- Proschan, M. A., and Hunsberger, S. A. (1995). Designed extension of studies based on conditional power. *Biometrics*, **51**, 1315-1324.
- Proschan, M. A., and Wittes, J. (2000). An improved double sampling procedure based on the variance. *Biometrics*, **56**, 1183-1187.

- SAS Institute (2004) *SAS/IML*[®] 9.1 *User's Guide, Volumes 1 and 2*. Cary, NC: SAS Institute.
- Shao, J. and Feng, H. (2007). Group sequential t -test for clinical trials with small sample sizes across stages. *Contemporary Clinical Trials*, **28**, 563–571.
- Shih, W. J. (2006). Group sequential, sample size re-estimation and two-stage adaptive designs in clinical trials: A comparison. *Statistics in Medicine*, **25**, 933-941.
- Shih, W. J., and Gould, A. L. (1995). Reevaluating design specifications of longitudinal clinical-trials without unblinding when the key response is rate of change. *Statistics in Medicine*, **14**, 2239-2248.
- Spiegelhalter, D. J., Freedman, L. S., and Blackburn, P. R. (1986). Monitoring clinical-trials - conditional or predictive power. *Controlled Clinical Trials*, **7**, 8-17.
- Spurrer, J. D. (1982). 2-stage tests of hypotheses in the general linear-model. *Communications in Statistics Part A-Theory and Methods*, **11**, 2775-2791.
- Stein, C. (1945), A Two-Sample Test for a Linear Hypothesis Whose Power is Independent of the Variance. *Annals of Mathematical Statistics*, **16**, 43-58.
- Totonchi and Guyuron (2007). A randomized, controlled comparison between arnica and steroids in the management of postrhinoplasty ecchymosis and edema. *Plastic and Reconstructive Surgery*, **120**: 271-274.
- Tsiatis A.A. (2006). Information-based monitoring of clinical trials. *Statistics in Medicine*, **25** : 3236-3244
- Tsiatis, A.A., and Mehta, C. (2003). On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika*, **90**, 367-378.
- Wald, A. (1947). *Sequential Analysis*, New York: Wiley.
- Wang, S.K., and Tsiatis, A. A. (1987). Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics*, **43**, 193-199.
- Wassmer, G. (1998). A comparison of two methods for adaptive interim analyses in clinical trials. *Biometrics*, **54**, 696-705.
- Whitehead, J. (1997). *The Design and Analysis of Sequential Clinical Trials*, revised 2nd Edition, Chichester: Wiley.
- Whitehead, J., and Stratton, I. (1983). Group sequential clinical-trials with triangular continuation regions. *Biometrics*, **39**, 227-236.
- Whitehead, J., Whitehead, A., Todd, S., Bolland, K., and Sooriyarachchi, M. R. (2001). Mid-trial design reviews for sequential clinical trials. *Statistics in Medicine*, **20**, 165-176.

Wittes, J., and Brittain, E. (1990). The role of internal pilot-studies in increasing the efficiency of clinical-trials. *Statistics in Medicine*, **9**, 65-72.

Zucker, D. M., and Denne, J. (2002). Sample-size redetermination for repeated measures studies. *Biometrics*, **58**, 548-559.

Zucker, D. M., Wittes, J. T., Schabenberger, O., and Brittain, E. (1999). Internal pilot studies II: Comparison of various procedures. *Statistics in Medicine*, **18**, 3493-3509.